Learning dynamics and decision paradigms in social-ecological dilemmas

Modeling complex social-ecological systems
based on the agent-environment interface

D I S S E R T A T I O N

zur Erlangung des akademischen Grades

d o c t o r r e r u m n a t u r a l i u m
(Dr. rer. nat.)
im Fach Physik
Spezialisierung: Theoretische Physik

eingereicht an der
Mathematisch-Naturwissenschaftlichen Fakultät
der Humboldt-Universität zu Berlin

von
M. Sc. Wolfram Martin Barfuß

Präsidentin der Humboldt-Universität zu Berlin:
Prof. Dr.-Ing. Dr. Sabine Kunst

Dekan der Mathematisch-Naturwissenschaftlichen Fakultät:
Prof. Dr. Elmar Kulke

Gutachter:

1. Prof. Dr. Dr. h.c. mult. Jürgen Kurths
2. Prof. Dr. Matjaž Perc
3. Dr. Konstantin Klemm

Tag der mündlichen Prüfung: 25. April 2019

Für Papa, Peter und Puma

Abstract

Collective action is required to enter sustainable development pathways in coupled social-ecological systems, safely away from dangerous tipping elements. Yet, in order to investigate the preconditions for cooperation, there is the challenge how to formally understand such social-ecological systems from a conceptual, mathematical modelers perspective.

Without denying the usefulness of other model design principles, this thesis proposes the agent-environment interface as the mathematical foundation for the design of social-ecological system models. From this perspective, it extends the concept of a social dilemma to a social-ecological dilemma. Social dilemmas have often been studied by evolutionary dynamics in repeated games with only one environmental state. Instead, this thesis uses stochastic games with multiple environmental states to investigate social-ecological dilemmas. Thereby, it extends the domain of social physics to social-ecological physics.

Yet, the majority of previously used evolutionary dynamics are not able to deal with multi-state environments. Therefore, this thesis focuses on the related concept of learning dynamics. It refines techniques from the statistical physics literature on learning dynamics to derive a deterministic limit of established reinforcement learning algorithms from artificial intelligence research. This enables a dynamical systems perspective on reinforcement learning in multi-state environments. Illustrations of the resulting learning dynamics of different learning algorithms across multiple example environments reveal a wide range of different dynamical regimes, such as fixed points, periodic orbits and deterministic chaos.

Eventually, this thesis applies the derived multi-state learning equations to a particular newly introduced environment, referred to as the Ecological Public Good. It models a coupled social-ecological dilemma, extending established repeated social dilemma games, such as the Prisoner's dilemma, by an ecological tipping element. The preconditions for both the emergence and stability of cooperation are investigated using a combination of numerical and analytical methods. This model is able to explain empirical observations as well as to reproduce known theoretical results. Novel qualitatively different parameter regimes are discovered, including one in which agents prefer to collectively suffer in environmental collapse rather than cooperating in a prosperous environment. Further, it can be shown that cooperation can remain stable despite considerable shortsightedness of the agents. However, this is only the case if the expected damage in the case of collapse is large. Conversely, this means that such reward-optimizing learners will break off the cooperation agreement, if they do not believe in likely and severe consequences of a tipping catastrophe.

Since optimization approaches have also been criticized in other contexts of environmental governance, this thesis challenges the reward optimizing paradigm of the learning equations. Prominent alternatives to the decision paradigm of economic optimization are sustainability and the safe operating space. This thesis presents a novel formal comparison of these three decision paradigms for the governance of an environmental tipping element. There, it can be shown that optimization alone can lead to safe and sustainable behavior and policies, but is by no means guaranteed to do so. In fact, no paradigm guarantees fulfilling requirements imposed by another paradigm. Further, the absence of a master paradigm is shown to be of special relevance for governing the climate system, since the latter may reside at the edge between parameter regimes where economic welfare optimization becomes neither sustainable nor safe.

In summary, this thesis demonstrates the usefulness of the agent-environment interface for the design of social-ecological system models, leading to a deeper theoretical understanding of such systems and in particular of the preconditions for successful collective action towards ecologically safe and socially just sustainability.

Zusammenfassung

Kollektives Handeln ist erforderlich um nachhaltige Entwicklungspfade in gekoppelten sozial-ökologischen Systemen zu erschließen, fernab von gefährlichen Kippelementen. Um die Voraussetzungen für die notwendige Kooperation zu erforschen, besteht jedoch aus der Sicht eines konzeptionellen, mathematischen Modellierers die Herausforderung, wie solche sozial-ökologischen Systeme formal zu verstehen sind.

Ohne anderen Modellierungsprinzipien ihren Nutzen abzuerkennen, schlägt diese Dissertation die Agent-Umwelt Schnittstelle als die mathematische Grundlage für das Modellieren sozial-ökologischer Systeme vor. Aus dieser Perspektive heraus erweitert sie das Konzept eines sozialen Dilemmas zu einem sozial-ökologischen Dilemma. Soziale Dilemmata werden oft mit Hilfe evolutionärer Dynamiken in sich wiederholenden Spielen mit nur einem Umweltzustand untersucht. Diese Arbeit dagegen verwendet stochastische Spiele, die mehrere Umweltzustände besitzen können, zur Untersuchung sozial-ökologischer Dilemmata. Damit erweitert sie den Bereich der Sozialphysik zu sozial-ökologischer Physik.

Jedoch ist die Mehrheit der bisher verwendeten evolutionären Dynamiken nicht in der Lage mit Umwelten umzugehen, die aus mehreren Zuständen bestehen. Daher konzentriert sich diese Arbeit auf das verwandte Konzept der Lerndynamiken. Sie erweitert eine Methode aus der Literatur der statistischen Physik über Lerndynamiken, um einen deterministischen Grenzübergang von etablierten Verstärkungslernalgorithmen aus der Forschung zu künstlicher Intelligenz herzuleiten. Diese erlauben eine dynamische Systemperspektive auf das Lernen in Umwelten mit mehreren Zuständen. Die resultierenden Lerndynamiken verschiedener Algorithmen zeigen in mehreren Beispielumwelten eine große Bandbreite verschiedener dynamischer Regime wie z.B. Fixpunkte, Grenzzyklen oder deterministisches Chaos.

Schließlich werden die hergeleiteten Lerngleichungen auf eine bestimmte, neu eingeführte Umwelt angewendet, hiermit bezeichnet als Ökologisches Öffentliches Gut. Sie modelliert ein gekoppeltes sozial-ökologisches Dilemma und erweitert damit etablierte soziale Dilemmaspiele wie z.B. das Gefangenendilemma um ein ökologisches Kippelement. Es werden die Voraussetzungen sowohl für die Entstehung als auch für die Stabilität von Kooperation mit Hilfe einer Kombination aus numerischen und analytischen Methoden untersucht. Dieses Modell ist in der Lage sowohl empirische Beobachtungen zu erklären als auch zuvor bekannte theoretische Ergebnisse zu reproduzieren. In dieser Arbeit werden neuartige, qualitativ verschiedene Parameterregime aufgezeigt, darunter eines, in dem Agenten es vorziehen, gemeinsam unter einem Kollaps der Umwelt zu leiden, als in einer florierenden Umwelt zu kooperieren. Weiter kann gezeigt werden, dass Kooperation trotz beträchtlicher Kurzsichtigkeit der Agenten stabil bleiben kann. Dies ist jedoch nur dann der Fall, wenn der zu erwartende Schaden im Falle eines Kollapses groß ist. Umgekehrt bedeutet dies, dass belohnungsoptimierende Lern-Agenten die Kooperationsvereinbarung kündigen, wenn sie nicht an die Möglichkeit und die schwerwiegenden Folgen einer Kipp-Katastrophe glauben.

Da auch in anderen Zusammenhängen von Umweltmanagement Optimierungsansätze kritisiert wurden, stellt diese Arbeit das Optimierungsparadigma der Lern-Agenten in Frage. Prominente Alternativen zum Entscheidungsparadigma der ökonomischen Optimierung sind Nachhaltigkeit und der sichere Handlungsraum. Diese drei Paradigmen werden systematisch miteinander verglichen, während diese auf das Management eines umweltlichen Kippelements angewendet werden. Es kann gezeigt werden, dass Optimierung allein dazu in der Lage ist, zu sicherem und nachhaltigem Verhalten bzw. Politiken zu führen, dies jedoch keineswegs garantiert ist. Tatsächlich ist es so, dass kein Paradigma garantiert, Anforderungen anderer Paradigmen zu erfüllen. Darüber hinaus wird gezeigt, dass das Fehlen eines Meisterparadigmas

von besonderer Bedeutung für das Klimasystem ist, da dieses sich möglicherweise am Rand zwischen Parameterbereichen befindet, wo ökonomische Optimierung weder nachhaltig noch sicher wird.

Insgesamt zeigt diese Dissertation damit die Nützlichkeit des Konzepts der Agent-Umwelt Schnittstelle für die Modellierung sozial-ökologischer Systeme. Dies führt zu einem tieferen theoretischen Verständnis solcher Systeme und insbesondere der Voraussetzungen für gelingendes, kollektives Handeln in Richtung ökologisch sicherer und sozial gerechter Nachhaltigkeit.

Acknowledgements

This thesis would not exist without the support of many individuals and organizations.

I thank the Potsdam Institute for Climate Impact Research (PIK) for hosting me and providing such a unique environment for scientific research on agent-environment systems. I thank Prof. Jürgen Kurths for his continuous support throughout the last three and half years. I thank all referees for their willingness to evaluate this thesis. I am deeply grateful for Jonathan Donges' day-to-day supervision, for allowing me the agency to follow my scientific curiosity and for our countless discussions. Especially our joint travels to Stockholm, Oxford, Princeton and other places were always sources of inspiration.

This work flourished in the context of the COPAN project on Coevolutionary Pathways in the Earth system at PIK. Thanks go to Finn Müller-Hansen, Ilona Otto, Jakob Kolb, Jobst Heitzig, Jonathan Donges, Marc Wiedermann, Nico Wunderling, Tim Kittel, Wolfgang Lucht and many many more for creating these flourishing grounds. I thank Avit Bhowmik, Caroline Schill, Emilie Lindkvist, Ingo Fetzer, Jon Norberg, Maja Schlüter, Sarah Cornell, Steven Lade, Tiina Häyhä, and many more for making my stays at the Stockholm Resilience Centre (SRC) always an enriching experience.

I am grateful for the trust the Heinrich-Böll Foundation (HBF) has placed in me and the early stages of this project. Its scholarship was much more than just financial support. I am thankful for the truly multidisciplinary experiences and especially the friendships I have found within the Böll network. I thank the Wilhelm and Else Heraeus Foundation for funding my participation at the DPG conferences. And I thank the Princeton-Humboldt Cooperation and Collective Cognition Network (CoCCoN), and there within especially Pawel Romanczuk, for the opportunity to participate and providing the funds to visit Princeton University.

I thank all members of PIK, SRC and HBF who supported me with organizational assistance and technical infrastructure. I gratefully acknowledge the European Regional Development Fund (ERDF), the German Federal Ministry of Education and Research and the Land Brandenburg for supporting my work by providing resources on the high performance computer system at PIK. And I am thankful to all developers who provide Linux operating systems and software and the python language as open-source software.

I thank all my co-authors of the publications underlying this thesis. Furthermore, I thank Amelie Reinisch, Felix Strnad, Finn Müller-Hansen, Helmut Barfuss, Jakob Kolb, Jonathan Donges, Jürgen Kurths, Marc Wiedermann and Simon Levin for comments on parts of this manuscript. All remaining errors are mine.

Finally, my thanks go to my parents and friends. I can always count on you.

Contents

List of publications

This dissertation is partly based on the following publications. The identifiers (e.g. P1), given below are cited in the text to highlight passages that are connected to one or more of these papers.

P8 (in preparation) **W. Barfuss**, J.F Donges, J. Kurths, S. Levin. *On the emergence and stability of cooperation in the ecological public good.*

P7 **W. Barfuss**, J.F Donges, J. Kurths (2018). *Deterministic limit of temporal difference reinforcement learning for stochastic games.* Physical Review E 99(4), 043305

P6 (in review) J.F. Donges, W. Lucht, J. Heitzig, **W. Barfuss**, S.E. Cornell, S.J. Lade, M. Schlüter (2018), *Taxonomies for structuring models for World-Earth system analysis of the Anthropocene: subsystems, their interactions and social-ecological feedback loops.* Earth System Dynamics Discussions

P5 (in review) J.F. Donges, J. Heitzig, **W. Barfuss**, J.A. Kassel, T. Kittel, J.J. Kolb, T. Kolster, F. Müller-Hansen, I.M. Otto, M. Wiedermann, K.B. Zimmerer, W. Lucht (2018). *Earth system modeling with complex dynamic human societies: the copan:CORE World-Earth modeling framework.* Earth System Dynamics Discussions

P4 **W. Barfuss**, J.F. Donges, S.J. Lade and J. Kurths (2018). *When optimization for governing human-environment tipping elements is neither sustainable nor safe.* Nature Communications 9(1), 2354

P3 J. Heitzig, **W. Barfuss** and J.F. Donges (2018). *A thought experiment on sustainable management of the Earth system.* Sustainability 10(6), 1947

P2 J.F. Donges and **W. Barfuss** (2017). *From math to metaphors and back again. Social-ecological resilience from a multi-agent-environment perspective* GAIA 26(1), 182–190

P1 **W. Barfuss**, J.F. Donges, M. Wiedermann and W. Lucht (2017). *Sustainable use of renewable resources in a stylized social-ecological network model under heterogeneous resource distribution.* Earth System Dynamics 8(2), 255–264

List of figures

List of tables

List of frequently used mathematical symbols

N	Number of agents
i, j	Individual agents
t	Time
M	Number of actions
$\mathcal{A}$	Action set
$\mathcal{A}^i$	Agent i's action set
$\boldsymbol{\mathcal{A}}$	Joint action set
$\boldsymbol{\mathcal{A}}^{-i}$	Joint action set except agent i's
a	Action $a \in \mathcal{A}$ or $a \in \mathcal{A}^i$
$\mathbf{a}$	Joint action $\mathbf{a} \in \boldsymbol{\mathcal{A}}$
$\mathbf{a}^{-i}$	Joint action except agent i's $\mathbf{a}^{-i} \in \boldsymbol{\mathcal{A}}^{-i}$
c	The cooperation action
d	The defection action
l	The low pressure action
h	The high pressure action
Z	Number of environmental states
$\mathcal{S}$	Environmental state set
s	State $s \in \mathcal{S}$
s'	Next state $s' \in \mathcal{S}$
$\mathbf{T}$	Transition tensor
$T_{s\mathbf{a}s'}$	Entry of transition tensor
p	The prosperous state
g	The degraded state
$\mathbf{R}$	Reward tensor
$R^i_{s\mathbf{a}s'}$	Entry of reward tensor
$\mathbf{X}$	Behavior profile *or* Policy
X^i_{sa}	Entry of behavior profile *or* of policy
$\boldsymbol{\sigma}(\mathbf{X})$	Stationary state distribution
$V^i_s(\mathbf{X})$	State-value
$Q^i_{sa}(\mathbf{X})$	State-action-value
$\tilde{V}^i_s(t)$	State-value approximation
$\tilde{Q}^i_{sa}(t)$	State-action-value approximation
α	Learning rate of an agent
β	Exploitation level of an agent
γ	Farsightedness of an agent

Chapter 1

Introduction

That's the beauty of grand opera: you can do anything ...as long as you sing it.

Anna Russell

1.1 Sustainability in coupled social-ecological systems

Sustainability. Prospering within planetary boundaries is the human quest of the 21st century (Rockström and Klum, 2012). Planetary boundaries are critical thresholds of nine key Earth system processes that, if crossed, are likely to cause the Earth system to tip into a disastrous, unfavorable state for humanity (Rockström et al., 2009b). Two examples of such processes are climate change and the loss of biosphere integrity. Both have been associated with the two core planetary boundaries, each of which has the potential on its own to drive the Earth system into a new state, should they be substantially and persistently transgressed (Steffen et al., 2015a). Thus, these nine planetary boundaries constitute a *safe operating space* for humanity with respect to the Earth system (Rockström et al., 2009a).

It has been put forward that humanity must not only stay below planetary boundaries, but simultaneously above social foundations (Raworth, 2012). Examples of such social foundations include basic education, decent health or the ending of poverty (Raworth, 2017). They derive themselves from the Universal Declaration of Human Rights (UN General Assembly, 1948) which constitutes a global perspective of the minimum standards required for a prospering human life in dignity. Together, planetary boundaries and social foundations delineate *a safe and just operating space* for humanity (Raworth, 2012, 2017).

Collective human action is required to enter such a safe and just operating space (Griggs et al., 2013; Steffen et al., 2018). From a practical policy perspective, important declarations in this regard include the United Nations' Brundtland report (WCED, 1987), which sets the course towards sustainable development. Influenced by human rights and planetary boundaries concepts, the resolution of the 17 Sustainable Development Goals (SDGs, UN General Assembly, 2015) and the adoption of the Paris Agreement on climate change (COP, 2015) in the same year mark the world's nations agreement to cooperate towards a sustainable future.

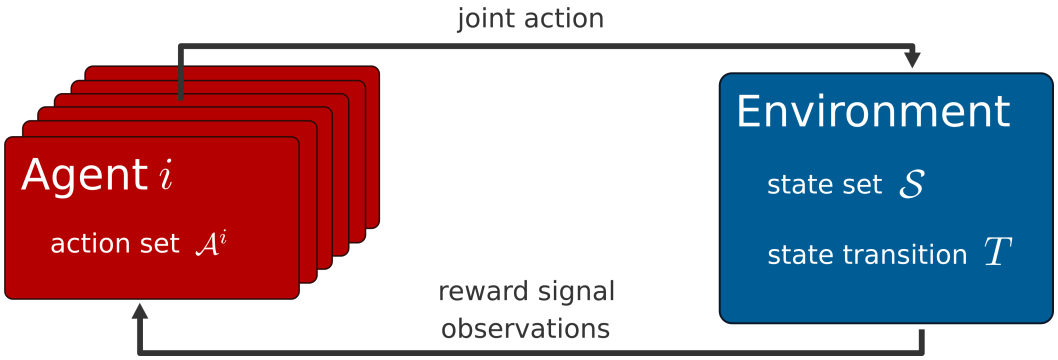

Figure 1.1: Agent-environment interface

However, it remains a political challenge whether countries will actually meet the set targets. Increasing political polarization (Dunlap et al., 2016) makes the question arise how stable these cooperation agreements actually are.

Social-ecological systems. For such collective efforts to cooperate towards a sustainable future it is important to not treat ecological or social systems as isolated, but acknowledge them as coupled complex adaptive social-ecological systems (Berkes and Folke, 1998; Levin et al., 2012). Especially in the last century, human activities have changed the relationship between societies and the environment up to the global scale (Crutzen, 2002; Steffen et al., 2007), making them mutually interdependent and the quest for understanding their joint dynamics urgent. Social-ecological system dynamics result from feedback loops involving human behavior, institutional and biophysical processes. Hence, these often non-linear dynamics involve tipping elements (Lenton et al., 2008; Schellnhuber, 2009), regime shifts (Lade et al., 2013; Scheffer et al., 2001), and multi-stabilities (Donges et al., 2017), as well as multiple kinds of uncertainties (Anderies et al., 2007; Irwin et al., 2016; Polasky et al., 2011a), and extreme events (Farmer et al., 2015).

These feedback effects bring into question what preconditions are required such that collective action towards sustainability can succeed. From a conceptual, mathematical modelers perspective, this raises the more fundamental question, how to formally understand social-ecological systems in order to investigate these preconditions for cooperation in social-ecological systems.

1.2 The agent-environment interface

The overall aim of this thesis is to propose the agent-environment interface as known from e.g. Sutton and Barto (1998) as a formal mathematical framework upon which social-ecological system models can be build. As such, it remains a proposal that has to prove its usefulness in the concrete applications dealt with in this thesis.

The key feature of the agent-environment interface (Fig. 1.1) is the clear separation of agents and environment. The concept of an *agent* is to be understood as an abstraction of any real-world entity that acts, i.e. chooses an action from its given action set. An agent can be a model of e.g. a human individual, a government, a firm, an organization, a city, etc., as the model designer desires. The *environment* is to be understood as the model parts that are not acting. Thus, it must not necessarily be composed of purely ecological model parts, as the name environment might suggest. Mathematically, it consists of a set of states, which may change probabilistically, depending on the agents' actions. Generally, agents are able to *observe* environmental states and other agents. They also receive a *reward* signal from the environment, which may depend on the joint action of the agents and the current state transition of the environment. Thus, agents generally gain rewards from the environment and from other agents. A more formal presentation of multi-agent environment systems based on the agent-environment interface principle will be given in Chapter 3.

Why does the agent-environment interface promise to be a useful tool for social-ecological systems modeling? The agent-environment interface offers a unifying perspective in the following sense: It is a mathematical design principle which occurs in many scientific fields. These occurrences allow fruitful connections between social-ecological systems research and those disciplines. Artificial intelligence research and reinforcement learning (Busoniu et al., 2008; Sutton and Barto, 1998), Markov decision processes (Puterman, 2005), optimal (Kirk, 2012), robust (Zhou and Doyle, 1998), viable (Aubin et al., 2011) control theory and cybernetics (Wiener, 1961), all use some form of an agent-environment interface, likewise does game theory, e.g. in evolutionary game theory (Perc et al., 2013; Perc and Szolnoki, 2010), the theory of learning in games (Fudenberg and Levine, 1998), and the theory of stochastic games (Neyman and Sorin, 2003; Shapley, 1953). Psychology and neuroscience are connected to the agent-environment interface via artificial intelligence research and reinforcement learning (Hassabis et al., 2017; Shah, 2012). Further, the majority of behavioral economic experiments is based on some form of a mathematical game and thereby also using an agent-environment interface (Camerer et al., 2004). The framework proposed by Schlüter et al. (2017) for mapping and comparing behavioral theories in models of social-ecological systems also uses an agent-environment interface, through which connections to social science theory can be drawn. Naturally, all these fields have developed their own unique perspective, including assumptions, terminology and notation, to serve their specific individual research questions. The interested reader is referred to the works of Shoham and Leyton-Brown (2008) and Gintis (2014) for more elaborate presentations towards a unified perspective.

The agent-environment interface is a simple and concrete design principle, upon which social-ecological system models can be build. Unlike less formalized theorizations of social-ecological systems e.g. by Bodin and Tengö (2012), Hinkel et al. (2014), and Ostrom (2007, 2009) social-ecological systems modeled through the agent-

environment interface can be easily put into mathematical practice. It thus gives one concrete answer to the question on how to mathematically model social-ecological systems (van Vuuren et al., 2016; Verburg et al., 2016). Nevertheless, it can be of great interest to make these less formalized social-ecological systems frameworks mathematically operationalizable as well (see e.g. Leslie et al., 2015).

Social-ecological systems research, in a broad sense, has already begun to utilize the agent-environment interface as its basic mathematical framework, e.g. with cybernetics (Schellnhuber, 1998), Markov decision processes (Chadès et al., 2016), robust control (Anderies et al., 2007; Rodriguez et al., 2010), viability theory (Heitzig et al., 2016; Martinet and Doyen, 2007), intelligent decision making and learning dynamics (Lindkvist and Norberg, 2014; Lindkvist et al., 2017; von der Osten, 2017) or behavioral experiments (Lindahl et al., 2016; Schill et al., 2015).

1.3 Overview and outline

From social to social-ecological physics. This thesis is a physicist's contribution to deepen the theoretical, conceptual understanding of coupled social-ecological systems by the means of mathematical models.

The idea of applying physics methods to social phenomena dates back centuries (Stauffer, 2012). While statistical physics applies the laws of statistics together with assumptions suitable for physics models, sociophysics uses similar methods for social models. Throwing a coin once, one cannot predict the side it falls upon. Asking one person how they voted cannot predict the outcome of an election. Yet, throwing many coins or asking many people can get one closer to an answer (Stauffer, 2012).

In particular, methods of statistical physics have proven to be valuable for studying the evolution of cooperation in social dilemma games (Perc et al., 2017). Here, especially the mechanism of network reciprocity (Nowak, 2006) has been a main driver for the involvement of statistical physics methods. Thus, it is well known that the structure of a network can significantly affect evolutionary outcomes (Perc and Szolnoki, 2010).

As the community is self-critically aware, efforts to become more empirically grounded is of crucial importance to establish the contribution of physicists to social dynamics (Castellano et al., 2009; Sánchez, 2018). Sánchez (2018) here suggests the equivalence of physics experiments to experiments of behavioral economics (Camerer et al., 2004). In those, a number of human subjects are asked to play the game of interest under various conditions. This kind of empirical research has already been influential regarding future directions for social physics (Capraro and Perc, 2018).

As Tavoni and Levin (2014) argue, a multidisciplinary approach is needed in order to tackle the increasingly pressing and intertwined environmental challenges faced by modern societies, this work is a physicist's contribution to this realm, thereby in turn contributing to the increasing multidisciplinarity of physics (Sinatra et al., 2015).

This thesis aims to deepen the conceptual understanding of social-ecological systems. Guided by the design principle of the agent-environment interface it extends

the concept of a social dilemma to a social-ecological dilemma. While social dilemmas have often been studied by evolutionary dynamics in repeated games, this thesis studies social-ecological dilemmas using so-called stochastic games. Stochastic games generalize repeated games by having an environment with multiple states, which can influence the agents' payoffs. This kind of extension results directly from the design principle of the agent-environment interface which makes the environment explicitly visible. Instead of evolutionary dynamics of populations this thesis focuses on learning dynamics of agents. However, both are structurally very similar, as argued in Chapter 4. Doing so, this thesis extends the domain of social physics to social-ecological physics.

Outline. This dissertation can be regarded as a story in three acts preceded by a prologue and a foundational chapter.

Chapter 2 is a prologue to the agent-environment interface. It presents a social learning network model of private renewable resource harvesting, which was not designed explicitly based on an agent-environment interface. Doing so, this chapter highlights the existence of different model design principles, of which all can be useful, depending on the questions asked. Yet, it may be difficult to reconcile them afterwards.

Chapter 3 introduces the conceptual and mathematical foundations of a special case of a multi-agent environment system, based on the agent-environment interface. The remainder of this thesis will utilize this special case, focusing on systems with finite action and state sets, as well as discrete update times.

The first act (Chapter 4) refines techniques from the statistical physics literature on learning dynamics to derive a deterministic limit of established reinforcement learning algorithms from artificial intelligence research. This is a necessary methodological extension, in order to enable a dynamical systems perspective on coupled social-ecological dilemmas in stochastic games with multiple environmental states, since the majority of previously used evolutionary dynamics are not able to deal with such environments. This chapter demonstrates the potential of this method with the three well established learning algorithms Q learning, SARSA learning and Actor-Critic learning. Illustrations of their dynamics on multi-agent, multi-state environments reveal a wide range of different dynamical regimes, such as convergence to fixed points, limit cycles and even deterministic chaos.

The second act (Chapter 5) applies the derived multi-state learning equations to a particular, newly introduced environment, termed the Ecological Public Good. It models a coupled social-ecological dilemma, extending established repeated social dilemma games, such as the Prisoner's dilemma, by an ecological tipping element. Both, the preconditions for the emergence and stability of cooperation are examined by a combination of numerical and analytical methods. This model is able to explain empirical observations and reproduce known theoretical results. Novel qualitatively different parameter regimes are discovered, among those one, in which agents prefer to collectively suffer in environmental collapse, rather than cooperating in a prosperous

environment. With respect to the stability of cooperation it can be shown that cooperation can remain stable despite considerable shortsightedness. However, this is only the case if the expected damage in the case of collapse is large. Conversely, these reward optimizing learners who do not believe in likely and severe consequences of a tipping catastrophe will break off the cooperation agreement.

The final act of this thesis (Chapter 6) challenges the reward optimizing paradigm of the learning equations, since also in other contexts of environmental governance optimization approaches have been criticized. Prominent alternatives to the decision paradigm of economic optimization are sustainability and the safe operating space. These three decision paradigms are systematically compared when applied for the management of a single-agent tipping element environment, which can be regarded as a single-agent version of the Ecological Public Good as introduced in Chapter 5. This chapter shows that optimization alone can lead to safe and sustainable behavior, but is by no means guaranteed to do so. In fact, no paradigm guarantees fulfilling requirements imposed by another paradigm. The absence of such a master paradigm is shown to be of special relevance for governing the climate system, which may reside at the edge between parameter regimes where economic welfare optimization becomes neither sustainable nor safe.

Chapter 7 concludes this thesis by summarizing its contributions and an outlook for future research.

Chapter 2

Prologue: Networked social learning of private renewable resource use

Der Menschenfreund im Bund mit aller Menschheit Feinde.
Hans Wurst - from Paul Dessau's *Einstein*: Prologue

This chapter examines the preconditions for sustainable harvesting of private renewable resources within a networked social learning model. Since this model was originally not designed above an agent-environment interface, it presents an interesting case to be compared to this design principle. While analyzing the presented network model, this chapter thereby highlights the existence of different model design principles for social-ecological systems modeling, of which all can be valuable. However, different model designs might be difficult to be reconciled in retrospect.

This chapter is primarily based on (Barfuss et al., 2017, P1). The comparison with the agent-environment interface (Sec. 2.4) is new material.

2.1 Introduction

Whether, when and how human usage of biophysical resources meets limits that produce feedbacks onto social functioning has a long history of controversial discussion (Malthus, 1798; Meadows et al., 1972; Rockström et al., 2009a). Especially in the last century, human activities have changed the relationship between nature and society at the global scale (Crutzen, 2002; Steffen et al., 2007, 2015b), making them mutually interdependent in an unprecedented manner and the question of their joint dynamics urgent. Social and ecological systems should therefore be studied not only in isolation but also as interlinked social-ecological systems (Berkes and Folke, 1998).

This chapter contributes to this debate by investigating properties of a stylized social system that cause the linked resource use system to either collapse or remain viable. Such a perspective also has important implications for the mathematical modeling of interdependent, global human-environment interactions (van Vuuren et al., 2016; Verburg et al., 2016). Typically, in present-day analysis the Earth system is either modeled from a purely biophysical point of view (Claussen et al., 2002) or

from a biophysical-economic one (van Vuuren et al., 2012), depending on the scope of the research question. However, both approaches do not take into account social dynamics beyond macroeconomic paradigms.

In this regard, the aim of this chapter is to conceptually explore avenues for a third strand of global modeling, next to the biophysical and biophysical-economic one, incorporating also social-cultural dynamics (c.f. Donges et al., 2018, P6). Rooted in a genuinely social-ecological perspective, the naming *World-Earth system models* emphasizes the free coevolution of the social and ecological components (Schellnhuber, 1998, 1999). While sophisticated models of this type are not yet available, the literature contains various modeling studies that incorporate potentially important features such as static interaction networks (Chung et al., 2013; Sugiarto et al., 2015) to depict stylized social dynamics (Auer et al., 2015; Holme and Newman, 2006), tele-coupling effects in a globalized society interacting through social networks (Bodin and Tengö, 2012; Janssen et al., 2006), social-ecological regime shifts (Lade et al., 2013; Scheffer et al., 2001) and (social) tipping elements (Bentley et al., 2014; Schellnhuber, 2009), structural re-organization occurring on adaptive social networks (Gross and Blasius, 2008; Sayama et al., 2013; Schleussner et al., 2016; Snijders et al., 2010) or structural transformations (Lade et al., 2017) and cultural preference dynamics due to traits such as imitation (Traulsen et al., 2010) or homophily (Centola et al., 2007; McPherson et al., 2001).

Section 2.2 sets out a mathematical model to demonstrate that social network interactions, imitation and homophily may have a profound influence on the environmental state, such as determining whether a collection of private renewable resources collapses from overuse or not. This chapter argues that more elaborate and sophisticated implementations of such social phenomena should receive attention in the future development of global system models, supplementing already established Earth system and integrated assessment models, neither of which at present include them.

As a particular case study for this chapter's model the effect of heterogeneously distributed resources is being investigated. This is important since in the real world actors do have access to different amounts of biophysical resources. This chapter examines under which combinations of parameters, characterizing a social learning network process, the model converges to a sustainable regime for different degrees of resource access heterogeneity. Parameters governing social learning dynamics are on the one hand a homophily parameter ϕ, addressing the propensity of nodes to establish interactions with nodes of the same kind (see Sec. 2.2 for a detailed model description). On the other hand, the timescale of social interaction τ quantifies the average time for social updates on the network. In this model, agents deliberately employ no form of individual learning with regard to the optimal harvesting strategy to emphasize the effects of the described social learning process. Individual learning will be explored in Chapters 4 and 5. Already for homogeneous resource access (Wiedermann et al., 2015) one observes a threshold in the parameter space of the model from non-sustainable to sustainable regimes at certain critical values ϕ_c and τ_c.

Since the concrete heterogeneous resource distribution is often unknown, the effect of its heterogeneity on the critical transition parameters ϕ_c and τ_c will be systematically investigated. It turns out that a heavy-tailed in comparison to a non-heavy-tailed resource distribution changes the model's behavior considerably. This is important as real-world resource data suggests that access to biophysical resources may indeed be distributed with heavy-tails.

Sec. 2.2 introduces the networked social learning model of private renewable resource harvesting and presents empirical data of heterogeneous resource access. While Sec. 2.3 discusses the results of this model study Sec. 2.4 discusses its model design with respect to the agent-environment interface. A summary closes this chapter in Sec. 2.5.

2.2 Model and methods

The model design is intended to conceptually explore the coevolution of socio-cultural with ecological dynamics and not to follow any specific real-world setting. On a conceptual level, human-environment interactions can happen in a common- or private-pools setting. Common-pool dilemmas have been studied extensively in the past (e.g. Hardin, 1968; Ostrom, 2015; Tavoni et al., 2012). Here, agents can retrieve information on other agent's harvesting strategy either via the ecological subsystem, i.e. the common-pool itself or via purely social interactions. In order to specifically focus on the latter of the two processes no transfer of information via the ecological system takes place. Thus, this model discards a common-pool setting in favor of individual and private resource stocks per agent. Wiedermann et al. (2015) introduced a model for such a setting for the special case of homogeneously distributed private resources. They showed distinct regimes in its parameter space, and provided analytical approximations of its dynamics. Here, this setting is refined for the more general case of an inhomogeneous resource distribution. Fig. 2.1 presents an overview of the model.

2.2.1 A stylized anthroposphere

The social learning (Bandura, 1977) process takes place in a network initialized as a random graph G (Erdös and Rényi, 1960) with nodes labeled by integer number $i = 1, \ldots, N$, representing social agents. It is based on two theoretical paradigms:

(i) Agents either change their strategy through *imitation* (Bahar et al., 2014; Traulsen et al., 2010) or

(ii) adapt their local network structure by rewiring to other nodes with similar behavior (*homophily,* Centola et al., 2007; McPherson et al., 2001).

In order to integrate this discrete update process (Holme and Newman, 2006; Zanette and Gil, 2006) with the continuous evolution of the resource stocks, *social*

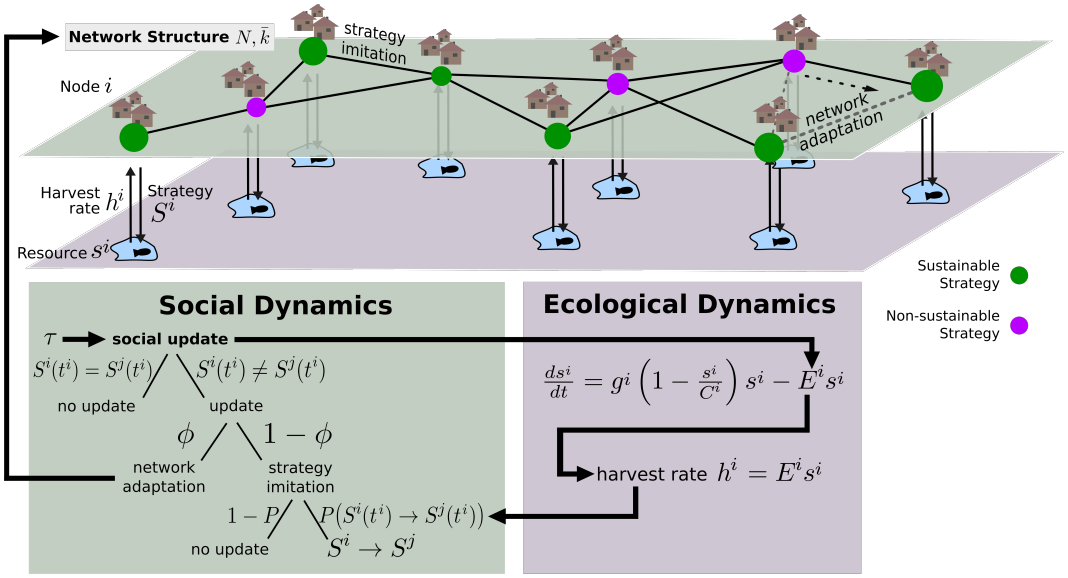

Figure 2.1: Model illustration. As the ecological sub-process the agents harvest their private logistically growing renewable resource with either a sustainable (green) or non-sustainable (purple) strategy. The social sub-process follows the logics of strategy imitation due to comparisons of harvest rates and of social network adaptation due to homophily. The social update times are generated by a Poisson process with average inter-event time τ.

update times t^i are assigned to the agents as generated by a Poisson process with an exponential distribution

$$p(\Delta t^i;\tau) = \frac{1}{\tau}\exp\left(\frac{-\Delta t^i}{\tau}\right) \tag{2.1}$$

of waiting times Δt^i, where the parameter τ gives the expected waiting time.

Thus, agent i with the lowest update time in the queue performs the social update process according to:

(1) if the degree of agent i is zero (i.e. i has no neighbors), do nothing, otherwise choose a neighbor j of i at random.

(2) If j and i employ the same harvesting strategy $S^i = S^j$ (either sustainable or non-sustainable, see below), do nothing. Otherwise,

(2.1) with rewiring probability ϕ disconnect j from i and connect i to a randomly chosen agent k that employs the same strategy.

(2.2) If (2.1) was not chosen, change the strategy of i to the one of j according to the sigmoidal imitation probability function

$$P(S^i \to S^j) = \frac{1}{2}\left(\tanh\left(\lambda\left[h^j(t) - h^i(t)\right]\right) + 1\right). \tag{2.2}$$

Hence, the greater the harvest rate h^j (see below) of agent j with respect to the harvest rate h^i of agent i, the more likely agent i is to change its strategy to the one of agent j. Agents only consider their current yields when formulating their next harvesting strategy. This assumption reflects the agent's limited knowledge of their own and their neighbors' ecosystems. The parameter λ controls the slope of the imitation probability function (Eq. (2.2)), i.e. for $\lambda \to \infty$ agent i would always imitate agent j's strategy if $h^j(t) > h^i(t)$, while for $\lambda \to 0$ the imitation probability tends to $1/2$ and is independent of the agents' harvest rates. Therefore, one can interpret λ as *imitation tendency* parameter. Traulsen et al. (2010) found this sigmoidal shape of imitation probability in a behavioral experiment.

(3) For the next update, another waiting time is drawn from the exponential distribution (Eq. 2.1) and added to the update time of agent i.

2.2.2 A stylized ecosphere

Private resource dynamics

The model's ecological module consists of private renewable resources each following a logistic growth function, which is chosen as one of the simplest and most commonly used models of renewable resource dynamics in a constrained environment (Brander and Taylor, 1998; Keeling, 2000; Perman et al., 2003). Additionally, a harvest rate $h^i = E^i s^i$ is subtracted from the rate of change of the resource stock s^i. E^i denotes the *effort* of agent i. Thus, the dynamics of the ith resource is given by

$$\frac{ds^i}{dt} = g^i \left(1 - \frac{s^i}{C^i}\right) s^i - E^i s^i. \tag{2.3}$$

Here, g^i denotes the growth rate and C^i the carrying capacity of the ith resource stock. The strategy S^i of agent i can either be sustainable ($S^i = 1$), resulting in an effort $E^i_s = \frac{g^i}{2}$. Otherwise S^i is non-sustainable ($S^i = 0$) with an effort $E^i_n = \frac{3g^i}{2}$. These efforts have been chosen such that the sustainable strategy coincides with the maximum sustainable yield, whereas the non-sustainable strategy leads to the full depletion of the resource stock and, consequently, no harvest at all in the long term. Note that E^i_n and E^i_s are symmetrically separated from the critical effort $E^i_c = g^i$. The latter is defined such that for positive efforts below E^i_c the resource stock converges to a non-zero stationary state, whereas for efforts above E^i_c the resource stock collapses and converges to zero. When in interplay with the social update process, Eq. 2.3 is used as its analytically derived definite integral, which circumvents the need of any numerical integration methods.

Resource heterogeneity

Heterogeneous access to resources is operationalized by randomly distributing the resource capacities C^i according to a prescribed probability density function. For this purpose, the log-normal distribution

$$\ln \mathcal{N}(C;\mu,\sigma) = \frac{1}{C\sigma\sqrt{2\pi}} \exp\left[-\frac{(\ln C - \mu)^2}{2\sigma^2}\right], \quad C > 0, \tag{2.4}$$

is used, with parameters μ and σ (not to be confused with the standard deviation of C). It derives from the normal distribution: a positive random variable is log-normally distributed if its logarithm is normally distributed. The log-normal distribution is therefore applicable for positive valued quantities and has a heavy tail. σ and μ are the standard deviation and the mean of the logarithmic variable $\ln C$, respectively. The log-normal distribution occurs in variables from many fields, including biological and economic attributes (Sachs, 1984).

Fig. 2.2 shows exemplary empirical distributions of three different types of resources to illustrate that real-world resource data can be qualitatively described by a log-normal distribution with least square fits revealing different σ parameters:

(i) biocapacity[1] per country ($\sigma = 1.42$) computed from the Ecological Footprint Network (Ewing et al., 2008) representing the capacity of ecosystems to regenerate what people extract,

(ii) total renewable water resources data[2] ($\sigma = 1.98$) characterizing the maximum yearly amount of water available to each country for the year 2012, and

(iii) forested land area[3] per country ($\sigma = 3.83$) for the year 1991.

Although the agreement between the log-normal distribution and the data is far from perfect, Fig. 2.2 supports the use of a log-normal model for resource heterogeneity in this chapter's stylized social-ecological system model.

This distribution is utilized to investigate how resource heterogeneity affects the behavior of the model in comparison to the frequently studied homogeneous case. To study only the effect of different resource distributions and keep the total amount of available resource stock constant the parameter μ was adjusted according to $\mu(\sigma) = -\sigma^2/2$, resulting in a fixed value for $\langle C \rangle = 1$, the mean of the carrying capacities.

[1] downloaded at `http://www.footprintnetwork.org/images/uploads/NFA_2010_Results.xls` on October 14, 2014

[2] downloaded at `http://www.fao.org/nr/water/aquastat/data/query/index.html?lang=en` on November 25, 2015

[3] downloaded at `http://faostat3.fao.org/download/R/RL/E` on November 24, 2015

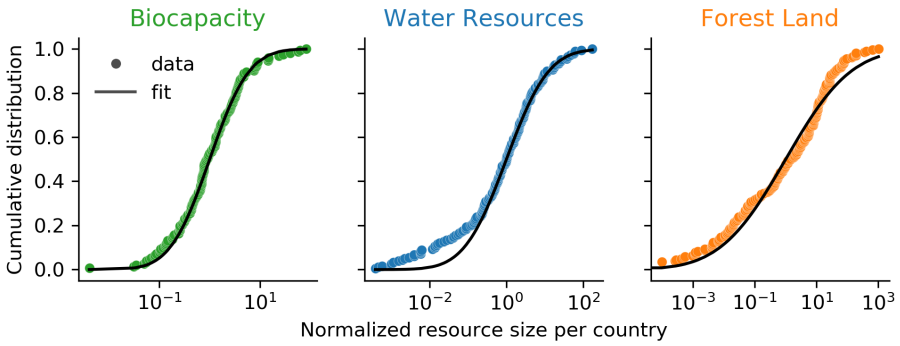

Figure 2.2: Empirical resource data per country normalized to the respective average (dots) together with least-square fitted log-normal distributions (lines): The biocapacity ($\sigma = 1.42$) computed from the Ecological Footprint Network (Ewing et al., 2008) represents the capacity of ecosystems to regenerate what people demand from them for the year 2007; the total renewable water resources ($\sigma = 1.98$) corresponds to the maximum theoretical yearly amount of water actually available for a country for the year 2012; forest land area per country ($\sigma = 3.83$) for the year 1991. The data are normalized to yield the same parameter $\mu = 0$ of the log-normal distribution. Note that the data qualitatively fits the log-normal distribution and that they give different values for the σ parameters of the log-normal distribution.

For comparison, results for non-heavy tailed resource capacities

$$C = |C_{\text{tmp}}|$$
$$\text{where} \quad C_{\text{tmp}} \sim \mathcal{N}(C_{\text{tmp}}; \mu_{\mathcal{N}}, \sigma_{\mathcal{N}}) = \frac{1}{\sigma_{\mathcal{N}}\sqrt{2\pi}} \exp\left[-\frac{(C_{\text{tmp}} - \mu_{\mathcal{N}})^2}{2\sigma_{\mathcal{N}}^2}\right] \tag{2.5}$$

are also presented, where $\mu_{\mathcal{N}}$ denotes the mean and $\sigma_{\mathcal{N}}$ the standard deviation of the underlying normal distribution. The resource heterogeneity $\sigma_{\mathcal{N}}$ is systematically varied on comparable ranges of variances for both - normal and log-normal - distributions, while the mean is also kept fixed ($\mu_{\mathcal{N}} = 1$). Since the normal distribution is not bounded by positive values, the absolute value of the drawn random variable is used.

2.2.3 Model parameterization and simulation protocol

A model run starts with an initial condition of stocks $s^i(0)$ uniformly distributed between 0 and C^i and harvesting strategies $S^i(0)$ drawn with a probability of 0.5 for a sustainable strategy $S^i = 1$ or a non-sustainable strategy $S^i = 0$. From the initial conditions, the model will converge to the *steady state* at t_f, where no further updates of strategy can occur. This is the case because the social network will consist solely of disconnected components with only one harvesting strategy (including the case of one single component) (Wiedermann et al., 2015). The remaining model parameters are the number of nodes $N = 500$, mean degree $\bar{k} = 20$, imitation tendency $\lambda = 1$ and ecological growth rate $g^i = 1$ for $i = 1, \ldots, N$ which are kept fixed throughout

the analysis. To account for the stochasticity inherent in the model, $R = 250$ runs for each parameter setting of interest were performed. The main observable is the fraction of sustainable harvesting nodes at the steady state

$$\langle S(t_f)\rangle_{N,R} = \left\langle \frac{1}{N}\sum_{i=1}^{N} S^i(t_f) \right\rangle_R \tag{2.6}$$

averaged over all ensemble runs R. $\langle S(t_f)\rangle_{N,R}$ is bounded between one and zero where $\langle S(t_f)\rangle_{N,R} = 1(0)$ denotes a completely (non-)sustainable regime.

2.3 Discussion of results

2.3.1 Social interaction timescale–homophily parameter space

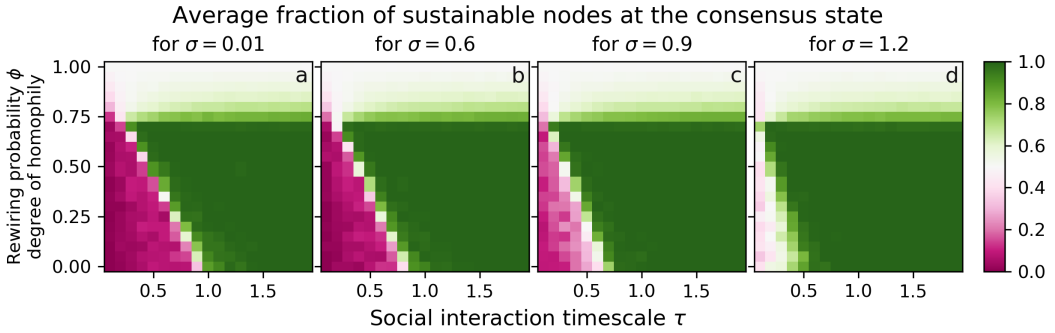

Figure 2.3: Social interaction timescale–homophily parameter space. Average fraction of sustainable harvesting agents in the steady state depending on the social network rewiring probability ϕ (measuring the degree of homophily) and the social interaction timescale τ for four distinct levels of resource heterogeneity: (a) $\sigma = 0.01$; (b) $\sigma = 0.6$; (c) $\sigma = 0.9$; (d) $\sigma = 1.2$. One observes four qualitatively different regimes: i) the sustainable regime for $\phi \lesssim 0.8$ and sufficiently large (slow) τ in green, ii) the non-sustainable or collapse regime for $\phi \lesssim 0.8$ and sufficiently small (fast) τ in purple, iii) in between both the transition regime in white as well as iv) the network fragmentation regime for $\phi \gtrsim 0.8$.

First, the effect of the rewiring probability ϕ (as a measure of the degree of homophily) and the average social interaction timescale τ on the fraction of sustainable harvesting nodes at the steady state $\langle S(t_f)\rangle_{N,R}$ (Eq. (2.6)) is investigated for vanishing resource heterogeneity ($\sigma = 0.01$) (Fig. 2.3 a).

Four regimes. Four qualitatively different regimes can be observed: the sustainable regime in green, the non-sustainable or collapse regime in purple, in between the transition regime in white and the network fragmentation regime for sufficiently large ϕ. The latter occurs since for large ϕ, social dynamics are dominated by homophily and, hence, by the process of social network rewiring, and thus negligibly few changes in strategy occur. The steady state is reached by a fragmentation of the network into

at least one purely sustainable and at least one purely non-sustainable component of comparable size. In turn, for smaller ϕ the effect of homophily is sufficiently weak such that most agents remain connected to a single component in the social network. The steady state is reached with a big connected network component. Here, large interaction timescales τ lead to a sustainable regime. This is because the comparisons of harvest rates typically happen when the logistic resource has been harvested for a sufficiently long time to reveal that the harvest rate converges to a positive value for a sustainable strategy whereas for a non-sustainable strategy it converges to zero.

Timescales. With respect to the timescale of social updates τ, it has been suggested that modern lifestyles are dominated by a social acceleration (Rosa, 2013). Simultaneously the pressure humanity is putting on the planet (Steffen et al., 2005) has experienced a great acceleration (Steffen et al., 2015b), threatening the Earth's ecosystems. Thus, a faster social timescales τ may lead to a non-sustainable regime, as observed in our model (see Fig. 2.3). Viewed with caution, the mechanisms in our model might be a possible explanation of this phenomenon. In any case, it highlights the importance of well synchronized social with ecological timescales. Since ecological timescales (e.g. the seasonal cycle) are difficult to influence, this suggests to take social timescales (e.g. election cycles, fashion trends, product launches) into account for possible policy interventions. As such it might be worthwhile to study the relationship between social and ecological timescales more intensively to identify suitable policy actions for the benefit of a sustainable system.

Homophily. Further, one can observe a linear relationship between critical parameters ϕ_c and τ_c where the transition between collapse and sustainable regimes occurs (Fig. 2.3). This result can be explained by the rate at which strategy changes happen. For $\phi = 0$, the transition occurs at $1/\tau \approx 1$, i.e. the ecological growth rate. For $\phi > 0$, imitation interactions happen at a rate $(1 - \phi)/\tau$ (Wiedermann et al., 2015) since the network rewires with probability ϕ and, hence, imitation takes place with probability $1 - \phi$. Hence, the effective imitation rate $(1 - \phi)/\tau$ equals approx. 1 (the ecological growth rate) in the transition regime, which explains the linear dependence between the two social parameters.

In other words, the homophily process in our model is beneficial for reaching the sustainable regime, where all agents harvest their resource gaining the maximum sustainable yield. All stochasticity and inherent shocks towards this sustainable steady state are absorbed. In this sense the sustainable regime can be described as resilient. This aligns with previous findings from Newig et al. (2010), who hypothesize that homophily has a beneficial effect on the resilience of a social-ecological network. Furthermore, one can interpret a large homophily parameter ϕ as the agent's means to protect themselves against the fast and free exchange of harvesting strategies. Along similar lines, it has been found that individuals with more environmental concerns hold also more protectionist policy preferences (Bechtel et al., 2012). This model suggests one possible mechanism how these relationships might come into

place. However, it remains to remark that a too large rewiring probability leads to a fragmentation of the social network into smaller groups of disjoint strategies, preventing the opportunity of a completely sustainable outcome. Thus, network adaptation at very high rates should be avoided for the sake of knowledge exchange and sustainable consensus formation.

Overall, these results demonstrate that immaterial processes distinct from macroeconomic optimization paradigms and residing exclusively in the social sphere, such as homophily and imitation, are capable of determining the eventual state of a material renewable resource. Thereby, these processes are able to govern a coupled social-ecological system such that full sustainability and total collapse are possible outcomes within the investigated social parameter space. Additionally, they show how the interaction of different social processes such as strategy imitation and homophily is able to shape the sustainable regime. This suggests that social-cultural processes should be considered as a potentially important part of feedback loops also in more elaborate models of the World-Earth system.

2.3.2 Systematic analysis of resource heterogeneity

Next, the effect of the resource heterogeneity σ on the transition regime between sustainable and non-sustainable steady states is investigated. The panels in Fig. 2.3 a-d show a qualitatively similar structure of parameter space for varying degrees of resource heterogeneity, yet with decreasing extent of the non-sustainable regime for increasing σ.

A more systematic analysis examines the average fraction of sustainable harvesting nodes at the consensus state $\langle S(t_f)\rangle_{N,R}$ for several segments of the parameter space spanned by τ, ϕ and the resource heterogeneity parameters σ ($\sigma_{\mathcal{N}}$), i.e. results are shown for both log-normally and normally distributed resource carrying capacities (Fig. 2.4). The ranges of σ for the log-normal and $\sigma_{\mathcal{N}}$ for the normal distribution are chosen such that they correspond to comparable standard deviations.

This analysis allows to explicitly show the effect of resource heterogeneity on the critical values τ_c (Fig. 2.4 a,c) and ϕ_c (Fig. 2.4 b,d), where the transition between the non-sustainable to the sustainable regime occurs. In general, the larger σ ($\sigma_{\mathcal{N}}$) the smaller τ_c and ϕ_c. In other words, a sustainable steady state can be achieved for faster social interactions and smaller degrees of homophily, the larger the resource heterogeneity is. The critical effective update timescale $\tau/(1-\phi) \stackrel{!}{=} \tau_{\text{eff,crit}}$ decreases to faster update times. This behavior is more pronounced for the log-normal distribution (Fig. 2.4 a,b) than for the normal one (Fig. 2.4 c,d) and can be explained by the heavy tails of the log-normal distribution. For a sufficiently large resource heterogeneity σ there is a sufficiently high probability that some agents will be assigned a comparably large resource capacity. Non-sustainable harvesting agents exploit their resources exponentially fast in time, whereas sustainable harvesting agents with comparably large resource capacity can retain their resource stock at a level that is still sufficiently large to convince other agents to become sustainable as well.

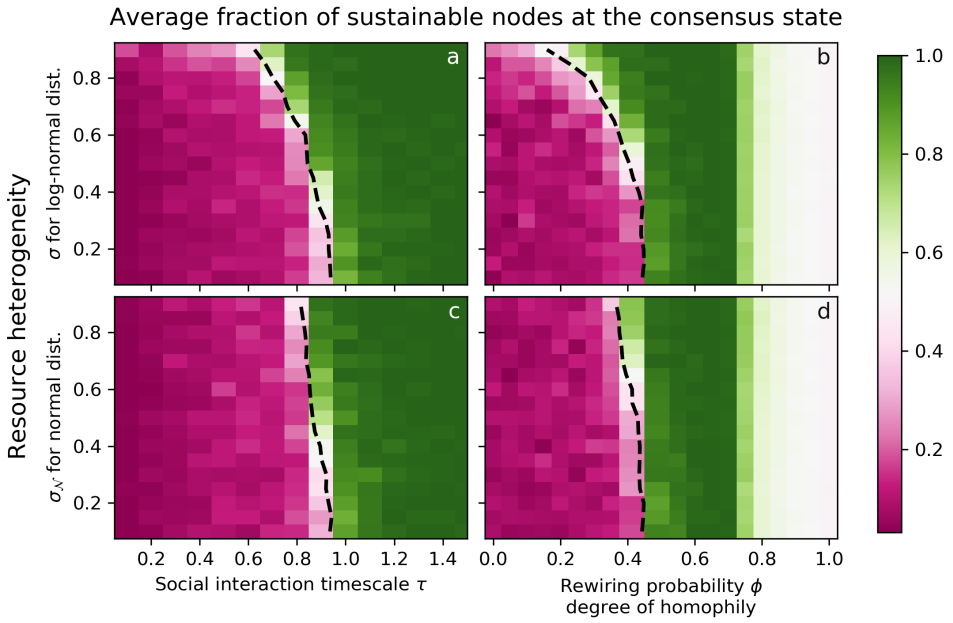

Figure 2.4: Effects of resource heterogeneity. Average fraction of sustainable harvesting nodes at the steady state for several segments of parameter space: (a,b) for (heavy-tailed) log-normally distributed capacities, and (c,d) for (non-heavy-tailed) normally distributed capacities. Parameter spaces spanned by social interaction timescale τ and resource heterogeneity σ ($\sigma_{\mathcal{N}}$) for rewiring probability $\phi = 0$ (a,c), and by ϕ and σ ($\sigma_{\mathcal{N}}$) for $\tau = 0.5$ (b,d). The ranges of σ and $\sigma_{\mathcal{N}}$ were chosen such that the standard deviations of both distributions are comparable. For both distributions, the mean was fixed to 1. The dashed black lines indicate the linearly interpolated 50% average fraction of sustainable nodes. Note the considerable effect the log-normal resource capacity distribution (in comparison to the normal distribution) has on the critical values of τ and ϕ, where the transition between the sustainable and the non-sustainable regime occurs.

At first, the observation that heterogeneity in the access to the private resources is enlarging the sustainable regime might be contradictory to one's intuition. This demonstrates the value of a thorough system's analysis and being critical about the own intuition. Cautiously comparing this phenomenon with the real-world one can interpret the size of the resource capacity as the effective economic power of international macro-agents, such as world-regions or nation states. This is justified, since no other economic processes are modeled but resource extraction; such as trade, innovation, labor, etc. The agents with comparably large economic power that employ a sustainable strategy have greater persuasive power than sustainable agents with smaller economic power. A country's energy transition to sustainable energy supply and its perceived impact on other countries might be a real-world example where a comparably strong economic country can exert also comparable large persuasive power to other countries to move forward towards sustainable energy supply.

Overall, heterogeneity to resource access in this model demonstrates, how comparably few sustainable first-movers with a large resource capacity are able to shift the overall system toward a sustainable state also at fast social interaction rates.

2.4 Discussion of model design

The design of this chapter's model represents an interesting case to be discussed with respect to the agent-environment interface. Although one can talk about agents in this model, it was not designed above such an interface. Such model designs follow a clear separation between agents (with actions, observations and rewards) and the environment (with states and dynamics between states). See Sec.1.2 for a more extensive discussion of the agent-environment interface and its advantages.

Nevertheless, one can reconstruct an agent-environment interface for this chapter's model in retrospect. In fact, various possibilities are conceivable (see below), and hence no *clear* agent-environment interface exists.

To reconstruct an agent-environment interface, the agents' actions and observations, as well as the rewards and environmental states have to be identified in line with the model's dynamics. To illustrate that various interfaces exist, highlighting possible choices of only the agents' action set is sufficient. One choice could be

$$\mathcal{A} = \{\text{rewire},\ \text{change harvesting strategy},\ \text{do nothing}\}.$$

Here, agents can only choose to either rewire, change their harvesting strategy or do neither of both. With this action set, the current harvesting strategy becomes part of the environmental state space. The agents have agency to either change or stay with their current harvesting strategy.

One might object, why should the agents not be able to simultaneously change their harvesting strategy and rewire. This combination would not be possible with the action set above. Yet, with an action set

$$\begin{aligned}\mathcal{A} &= \{\text{rewire},\ \text{not rewire}\} \times \{\text{harvest with low effort},\ \text{harvest with high effort}\} \\ &= \{\text{rewire and harvest with low effort},\ \text{rewire and harvest with high effort}, \\ &\quad \text{not rewire and harvest with low effort}, \\ &\quad \text{not rewire and harvest with high effort}\}\end{aligned}$$

simultaneously rewiring and changing the harvesting strategy would be possible. The circumstance that agents in this chapter's model in fact do not do both action processes simultaneously can be reflected in corresponding agent dynamics and does not mean that the underlying action set does not offer this choice. Additionally, this action set differs from the first one with respect to the harvesting strategies. Here, sustainable and non-sustainable harvesting are direct action processes. In the first example, the harvesting strategy was part of the environmental state set and the action was to change these states.

For both action set choices, if an agent chooses to rewire, the environment redistributes the agent to a new neighbor. Agents have no agency to which other agent they form a connection. This means, that the model assumption of homophily is no choice by the agents, but happens exclusively through the environment.

If the model designer wants to give the agents a choice to which other agent they form a connection,

$$\begin{aligned}\mathcal{A} = \{&\text{rewire from neighbor } j \text{ to node } k \mid \text{j} \in \text{agent's neighbors}, k \in \{1, \ldots, N\},\\ &\text{not rewire}\}\\ \times \{&\text{harvest with low effort}, \text{ harvest with high effort}\}.\end{aligned}$$

is a possible action set. If agents rewire, they choose which tie they break and which new tie they form. However, to reconcile the model's social dynamics, this action set would require that agents are able to observe the harvesting strategy of all other agents. These were only three possibilities. Further variants of the underlying agents' action sets could be easily constructed.

Model designs according to the agent-environment interface lead to a focus on individual agent behaviors and their emergent properties, resulting from the interplay between other agents and the environment. The advantage of the agent-environment interface design is that agent techniques, such a reinforcement learning (Chapter 4) or decision paradigms (Chapter 6) can be generally defined without a concrete environment in mind. However, these techniques require a distinct action set.

Designs such a the one of this chapter's model on the other hand, can be viewed from the perspective of social processes. Within the model, these social processes happen without the need to result from individual actions. Thus, this perspective does not require a clear separation between agents and environments.

Overall, it is important to acknowledge that both model design principles exist and that both are useful for answering research questions. Yet, it may be difficult to reconcile these both approaches after models have been constructed.

2.5 Summary

This chapter has investigated how social-ecological thresholds between sustainable and non-sustainable resource-use regimes depend on networked social learning interactions under conditions of resource heterogeneity. It used a stylized model of networked agents harvesting private renewable resources with either a sustainable or non-sustainable strategy. The strategies employed by the agents are updated through a social learning process on an adaptive social network reflecting an interconnected society. Resource heterogeneity was operationalized by log-normally and normally distributed carrying capacities of the resources.

It was shown that the properties of social processes such as strategy formation through imitation and homophilic social network adaptation alone can precondition the long-term state of renewable resources with outcomes ranging from environmental

collapse to sustainability. This observation suggests that a purely economic rationale may neglect decisive processes when modeling coupled social-ecological systems. It suggests that more sophisticated models of global coupled human-environment systems need to consider socio-cultural feedbacks as well (see also Donges et al., 2018, P6).

Furthermore, it was demonstrated that resource heterogeneities are important model ingredients that may not be neglected, especially when resource distributions possess heavy tails. This is relevant because empirical observations suggest that access to biophysical resources may indeed follow heavy tailed distributions.

Outlook. Overall, this chapter highlights how socio-cultural (i.e. immaterial) dynamics and interactions can have a profound qualitative effect on physical (i.e. material) states of the environment and, consequently, that neither social processes nor resource heterogeneities should be neglected a priori in a more sophisticated modeling of the World-Earth system.

In the context of the ongoing debate on global change (Steffen et al., 2005) and the Anthropocene (Crutzen, 2002; Steffen et al., 2007, 2015b), such more advanced models of planetary social-ecological systems (World-Earth models) are needed for developing a deeper understanding of the dynamics and interrelations between planetary boundaries (Rockström et al., 2009a; Steffen et al., 2015a) and social foundations (Raworth, 2012) for guiding humanity to a desirable safe and just operating space. Donges et al. (2018, P5) make a concrete suggestion for such a World-Earth modeling framework.

With respect to model design principles, it is important to acknowledge that different model design principles for social-ecological systems modeling exist. Yet, models of different designs might be difficult to be reconciled in retrospect. As a model designer, it is therefore important, to be conscious about the model design from the beginning.

The next chapter will introduce a variant of the agent-environment interface. Relevant concepts in sufficient mathematical detail will be given for multi-agent environment systems with discrete action and state sets as well as discrete update times. This variant of the agent-environment interface will be used throughout the rest of this thesis.

Python code of this chapter's model and scripts for the reproduction of the figures is available at github: `http://doi.org/10.5281/zenodo.1493202`. For illustrative purposes, a netlogo-version can be downloaded as well: `https://doi.org/10.5281/zenodo.1494178`.

Chapter 3

Multi-agent environment foundations

Ich habe auch ein sogenanntes Hobby, zu deutsch einen persönlichen Stil.

Krokodil - from Paul Dessau's *Einstein*: Intermezzo I

This chapter introduces a particular variant of multi-agent environment systems following the principle of the agent-environment interface. All following chapters of this thesis will utilize this variant. It is characterized by discrete action and environmental state sets, as well as discrete time steps. The environment is Markov, i.e. transition probabilities only depend on the current state of the environment and the current actions of the agents. The environment is assumed to be fully observable, i.e. agents can observe the current true Markov state. Essentially, such a setting is also known as a Markov or stochastic game (Neyman and Sorin, 2003; Shapley, 1953)

While the content presented in this chapter is known and used in many disciplines, this chapter's purpose is to give a common background, introduction of concepts and notational conventions for the following chapters.

It is based on parts of (Barfuss et al., 2019, P7).

3.1 Agents and the environment

There are $N \in \mathbb{N}$ *agents*. The environment can exist in $Z \in \mathbb{N}$ *states* $\mathcal{S} = \{S_1, \dots, S_Z\}$. In each state each agent has $M \in \mathbb{N}$ available *actions* $\mathcal{A}^i = \{A^i_1, \dots, A^i_M\}$, $i = 1, \dots, N$ to choose from. Having an identical number of actions for all states and all agents is notational convenience, no significant restriction. A joint action of all agents is referred to by $\mathbf{a} \in \boldsymbol{\mathcal{A}} = \mathcal{A}^1 \times \cdots \times \mathcal{A}^N$, the joint action of all agents but agent i is denoted by $\mathbf{a}^{-i} \in \boldsymbol{\mathcal{A}}^{-i} = \mathcal{A}^1 \times \cdots \times \mathcal{A}^{i-1} \times \mathcal{A}^{i+1} \times \cdots \times \mathcal{A}^N$.

Environmental dynamics are given by the probabilities for state changes expressed as a transition tensor

$$\mathbf{T} \in [0,1]^{Z \times M \times \dots (N \text{ times}) \dots \times M \times Z}. \tag{3.1}$$

The entry $T_{s\mathbf{a}s'}$ denotes the probability $P(s'|s,\mathbf{a})$ that the environment transitions to state s' given the environment was in state s and the agents have chosen the joint action $\mathbf{a}$. Hence, for all $s, \mathbf{a}$, $\sum_{s'} T_{s\mathbf{a}s'} = 1$ must hold. The assumption that the next state only depends on the current state and joint action makes this system Markovian. In this thesis environments are restricted to be ergodic, without absorbing states (c.f. Hennes et al., 2010).

Rewards. Agent receive rewards, given by the reward tensor

$$\mathbf{R} \in \mathbb{R}^{N\times Z\times M\times\ldots(N \text{ times})\ldots\times M\times Z}. \tag{3.2}$$

The entry $R^i_{s\mathbf{a}s'}$ denotes the reward agent i receives when the environment transitions from state s to state s' under the joint action $\mathbf{a}$. Rewards are also called payoffs from a game theoretic perspective.

Behavior. Agents draw their actions from their behavior profile

$$\mathbf{X} \in [0,1]^{N\times Z\times M}. \tag{3.3}$$

The entry $X^i_{sa} = P(a \,|\, i, s)$ denotes the probability that agent i chooses action a in state s. Thus, for all i and all s, $\sum_a X^i_{sa} = 1$ must hold.

This thesis focuses on the case of independent agents, able to fully observe the current state of the environment. With correlated behavior (see e.g. Busoniu et al., 2008) and partially observable environments (Oliehoek, 2012; Spaan, 2012) one could extend this multi-agent environment systems framework to be even more general.

Note that the behavior profile is usually called policy from a machine learning perspective or behavioral strategy from a game theoretic perspective. Policies and strategies might suggest a deliberate choice by the agents, the term *behavior* intends to avoid in the case of learning agents (Chapters 4 and 5). For the case of environmental governance Chapter 6 will speak of *polices* instead of behavior.

Fig. 3.1 refines Fig. 1.1 from the Introduction, illustrating this particular multi-agent environment system.

3.2 Behavioral and environmental averages

To allow a systematic averaging over the current behavior profile $\mathbf{X}$ and the environmental transitions $\mathbf{T}$, the following notational convention is introduced. Averaging over the whole behavioral profile yields

$$\begin{aligned} {}_{\mathbf{X}}\langle\circ\rangle &:= \sum_{\mathbf{a}} \mathbf{X}_{s\mathbf{a}} \cdot \circ \\ &:= \sum_{a^1\in\mathcal{A}^1} \ldots \sum_{a^N\in\mathcal{A}^N} X^1_{sa^1}\cdots X^N_{sa^N} \cdot \circ. \end{aligned} \tag{3.4}$$

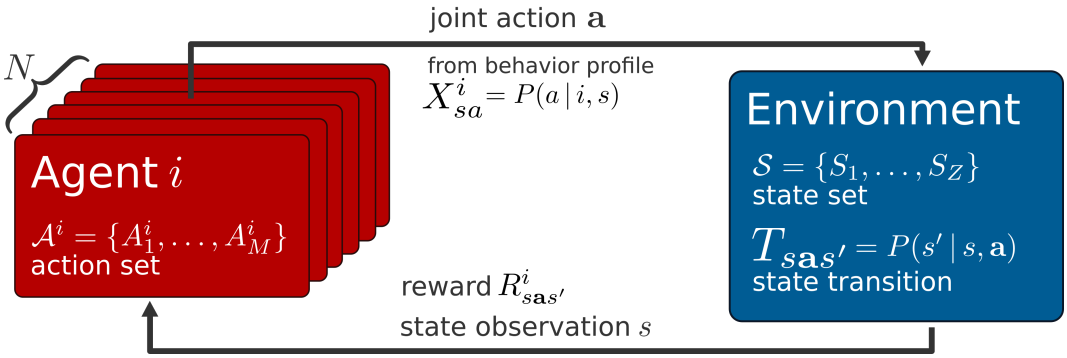

Figure 3.1: Multi-agent environment system. (also known as stochastic or Markov game). N agents choose a joint action $\mathbf{a} = (a^1, \ldots, a^N)$ from their action sets $\mathcal{A}^i$, based on the current state of the environment s, according to their behavior profile $X^i_{sa} = P(a|i,s)$. This will change the state of the environment from s to s' with probability $T_{s\mathbf{a}s'}$, and provide each agent with a reward $R^i_{s\mathbf{a}s'}$.

Here, $\circ$ serves as a placeholder. If the expression to be inserted for $\circ$ depends on the summation indices, then of course, those indices will be summed over as well. If the expression, which is averaged out, is used in tensor form, it is written in bold. If not, remaining indices are added after the right angle bracket.

Averaging over the behavioral profile of the other agents, keeping the action of agent i, yields

$$\begin{aligned} {}_{\mathbf{X}^{-i}}\langle \circ \rangle &:= \sum_{\mathbf{a}^{-i}} \mathbf{X}^{-i}_{s\mathbf{a}^{-i}} \cdot \circ \\ &:= \underbrace{\sum_{a^1 \in \mathcal{A}^1} \cdots \sum_{a^N \in \mathcal{A}^N}}_{\text{excl. } i} \underbrace{X^1_{sa^1} \cdots X^N_{sa^N}}_{\text{excl. } i} \cdot \circ \,. \end{aligned} \tag{3.5}$$

Last, averaging over the subsequent state s' yields

$${}_{\mathbf{T}}\langle \circ \rangle := \sum_{s'} T_{s\mathbf{a}s'} \cdot \circ := \sum_{s' \in \mathcal{S}} T_{sa^1 \ldots a^N s'} \cdot \circ. \tag{3.6}$$

Of course, these operations may also be combined as ${}_{\mathbf{TX}}\langle \circ \rangle$ and ${}_{\mathbf{TX}^{-i}}\langle \circ \rangle$, by multiplying both summations.

Effective Markov chain. For example, given a behavior profile $\mathbf{X}$, the resulting effective Markov Chain transition matrix reads ${}_{\mathbf{X}}\langle T \rangle_{ss'}$, which encodes the transition probabilities from state s to s'. From ${}_{\mathbf{X}}\langle T \rangle_{ss'}$ the stationary distribution of environmental states $\boldsymbol{\sigma}(\mathbf{X})$ can be computed. $\boldsymbol{\sigma}(\mathbf{X})$ is the row (or left) eigenvector corresponding to the eigenvalue 1 of ${}_{\mathbf{X}}\langle T \rangle_{ss'}$. Its entries encode the ratios of the average durations the agents find themselves in the respective environmental states.

Agent's rewards. The average reward agent i receives from state s under action a, given all other agents follow the behavior profile $\mathbf{X}$ reads ${}_{\mathbf{TX}^{-i}}\langle R\rangle^i_{sa}$. Including agent i's behavior profile gives the average reward it receives from state s: ${}_{\mathbf{TX}}\langle R\rangle^i_s$. Hence,

$$ {}_{\mathbf{TX}}\langle R\rangle^i_s = \sum_a X^i_{sa} \cdot {}_{\mathbf{TX}^{-i}}\langle R\rangle^i_{sa} \tag{3.7} $$

holds.

3.3 Agent's preferences and values

Return. Typically, agents are assumed to maximize their exponentially discounted sum of future rewards, called return

$$ G^i(t) = (1-\gamma^i)\sum_{k=0}^{\infty}(\gamma^i)^k r^i(t+k), \tag{3.8} $$

where $\gamma^i \in [0,1)$ is the *discount factor* or *farsightedness* of agent i and $r^i(t+k)$ denotes the reward received by agent i at time step $t+k$. Exponential discounting is most commonly used for its mathematical convenience and because it ensures consistent preferences over time. Other formulations of a return use e.g. finite time horizons, average reward settings, as well as other ways of discounting, such as hyperbolic discounting.

State-values. Given a behavior profile $\mathbf{X}$, the expected return defines the *state-value function*

$$ V^i_s(\mathbf{X}) := {}_{\mathbf{TX}}\Big\langle G^i(t) \mid s(t)=s\Big\rangle^i_s, \tag{3.9} $$

which is independent of time t, as one can see below. Here, the ${}_{\mathbf{TX}}\langle ...\rangle$ operator denotes the behavioral and environmental average as defined in Eqs. 3.4 and 3.6. Thus, the value of a state s for agent i is the expected return receivable from that state s, given that all agents behave according to $\mathbf{X}$

Combining Eq. 3.8 with Eq. 3.9 the well known Bellman equation (Bellman, 1957) can be derived as follows:

$$V_s^i(\mathbf{X}) = {}_{\mathbf{TX}}\Big\langle G^i(t) \mid s(t) = s \Big\rangle_s^i \tag{3.10}$$

$$= {}_{\mathbf{TX}}\Big\langle (1-\gamma^i) \sum_{k=0}^{\infty} (\gamma^i)^k r^i(t+k) \mid s(t) = s \Big\rangle_s^i \tag{3.11}$$

$$= {}_{\mathbf{TX}}\Big\langle (1-\gamma^i) r^i(t) + \gamma^i (1-\gamma^i) \sum_{k=0}^{\infty} (\gamma^i)^k r^i(t+1+k) \mid s(t) = s \Big\rangle_s^i \tag{3.12}$$

$$= {}_{\mathbf{TX}}\Big\langle (1-\gamma^i) r^i(t) + \gamma^i V_{s(t+1)}^i(\mathbf{X}) \mid s(t) = s \Big\rangle_s^i \tag{3.13}$$

$$= {}_{\mathbf{TX}}\Big\langle (1-\gamma^i) R_{s\mathbf{a}s'}^i + \gamma^i V_{s'}^i(\mathbf{X}) \Big\rangle_s^i . \tag{3.14}$$

Thus, this recursive relationship between state-values declares that the value of a state s is the discounted value of the subsequent state $s(t+1)$ plus $(1-\gamma^i)$ times the reward received along the way. For the reward $r^i(t)$ received at time step t the reward tensor $R_{s\mathbf{a}s'}^i$ can be inserted, resolving the state at time step t to be $s(t) = t$. The value of the subsequent state $V_{s(t+1)}^i(\mathbf{X})$ can be replaced with $V_{s'}^i(\mathbf{X})$.

Evaluating the behavioral and environmental average ${}_{\mathbf{TX}}\langle ... \rangle$ one can write:

$$V_s^i(\mathbf{X}) = (1-\gamma^i) \cdot {}_{\mathbf{TX}}\langle R \rangle_s^i + \gamma^i \sum_{s'} \cdot {}_{\mathbf{X}}\langle T \rangle_{ss'} \cdot V_{s'}^i(\mathbf{X}). \tag{3.15}$$

As above, the reward tensor average reads ${}_{\mathbf{TX}}\langle R \rangle_s^i$. The expected subsequent state-value $V_{s'}^i(\mathbf{X})$ can be written as a matrix multiplication of the effective Markov transition matrix and the vector of state-values: $\sum_{s'} {}_{\mathbf{X}}\langle T \rangle_{ss'} \cdot \mathbf{V}_{s'}^i(\mathbf{X})$.

Writing Eq. 3.15 in matrix form reads

$$\mathbf{V}^i(\mathbf{X}) = (1-\gamma^i) \cdot {}_{\mathbf{TX}}\langle \mathbf{R} \rangle^i + \gamma^i \cdot {}_{\mathbf{X}}\langle \mathbf{T} \rangle \cdot \mathbf{V}^i(\mathbf{X}). \tag{3.16}$$

A solution of the state-values $\mathbf{V}^i(\mathbf{X})$ can be obtained using matrix inversion

$$\mathbf{V}^i(\mathbf{X}) = (1-\gamma^i) \left(\mathbb{1}_Z - \gamma^i {}_{\mathbf{X}}\langle \mathbf{T} \rangle \right)^{-1} {}_{\mathbf{TX}}\langle \mathbf{R} \rangle^i . \tag{3.17}$$

The computational complexity of matrix inversion makes this solution strategy infeasible for large systems. Therefore many iterative solution methods exist (Wiering and Otterlo, 2012).

State-action-values. Equivalent to state-value functions, *state-action-value functions* Q_{sa}^i are defined as the expected return, given agent i applied action a in state s and then followed $\mathbf{X}$ accordingly:

$$Q_{sa}^i(\mathbf{X}) := {}_{\mathbf{TX}}\Big\langle G^i(t) \mid s(t) = s, a(t) = a \Big\rangle_{sa}^i . \tag{3.18}$$

They can be computed via

$$Q^i_{sa}(\mathbf{X}) = (1-\gamma^i)\,{}_{\mathbf{TX}^{-i}}\langle R\rangle^i_{sa} + \gamma^i \sum_{s'} {}_{\mathbf{X}}\langle T\rangle_{ss'} \cdot V^i_{s'}(\mathbf{X}). \tag{3.19}$$

Since the current reward results from the assumption, that the agent applied action a in state s the behavioral-environmental average ${}_{\mathbf{TX}^{-i}}\langle\ldots\rangle$ is used for the first term in Eq. 3.19. Afterwards the agent follows $\mathbf{X}^i$, resulting in the identical last term as in Eq. 3.15.

From Eq. 3.7 and the fact that $\sum_a X^i_{sa} = 1$ immediately follows that

$$V^i_s(\mathbf{X}) = \sum_a X^i_{sa} Q^i_{sa}(\mathbf{X}) \tag{3.20}$$

holds for the inverse relation of state-action- and state-values.

Value reward relation. Interestingly, one can show that the dot product between the stationary state distribution $\boldsymbol{\sigma}(\mathbf{X})$ of the effective Markov Chain with the transition matrix ${}_{\mathbf{X}}\langle\mathbf{T}\rangle$ and the behavior average reward ${}_{\mathbf{TX}}\langle\mathbf{R}\rangle^i$ is identical to the dot product of the stationary distribution and the state value $\mathbf{V}^i(\mathbf{X})$:

$$\boldsymbol{\sigma}(\mathbf{X}) \cdot \mathbf{V}^i(\mathbf{X}) = \boldsymbol{\sigma}(\mathbf{X}) \cdot {}_{\mathbf{TX}}\langle\mathbf{R}\rangle^i . \tag{3.21}$$

To show this relation Eq. 3.16 needs to be multiplied by $\boldsymbol{\sigma}(\mathbf{X})$:

$$\begin{aligned}
\boldsymbol{\sigma}(\mathbf{X}) \cdot \mathbf{V}^i(\mathbf{X}) &= (1-\gamma^i) \cdot \boldsymbol{\sigma}(\mathbf{X}) \cdot {}_{\mathbf{TX}}\langle\mathbf{R}\rangle^i + \gamma^i \cdot \boldsymbol{\sigma}(\mathbf{X}) \cdot {}_{\mathbf{X}}\langle\mathbf{T}\rangle \cdot \mathbf{V}^i(\mathbf{X}) && (3.22)\\
&= (1-\gamma^i) \cdot \boldsymbol{\sigma}(\mathbf{X}) \cdot {}_{\mathbf{TX}}\langle\mathbf{R}\rangle^i + \gamma^i \cdot \boldsymbol{\sigma}(\mathbf{X}) \cdot \mathbf{V}^i(\mathbf{X}) && (3.23)\\
&= \boldsymbol{\sigma}(\mathbf{X}) \cdot {}_{\mathbf{TX}}\langle\mathbf{R}\rangle^i , && (3.24)
\end{aligned}$$

using the fact that $\boldsymbol{\sigma}(\mathbf{X})$ is a left eigenvector of ${}_{\mathbf{X}}\langle\mathbf{T}\rangle$ and subsequently rearranging and dividing by the non-zero $(1-\gamma^i)$, since $\gamma^i \in [0,1)$.

In other words, the state-average value is identical to the state-average reward, independent of the discount factor γ^i. This equivalence qualifies this measure to express the performance of an agent in one single scalar.

3.4 Summary

This chapter introduced a variant of a multi-agent environment system, based the agent-environment interface principle. Basic concepts, such as the transition, reward and behavioral profile tensors, behavioral and environmental averages, an agent's return, state- and state-action values were introduced. These concepts will be put to use in the following chapters.

Chapter 4

First act: Deterministic limit of temporal difference reinforcement learning

Eine herrliche Nacht, die empörte Einbildungskraft zu verwildern.
Einstein - from Paul Dessau's *Einstein*: First act

While Chapter 2 focused on a social learning update process between networked agents, it left out any form of individual learning. This chapter brings the focus to individual learning through reinforcements, building up on the agent-environment framework introduced in Chapter 3. This chapter is necessary because, on the one hand, the majority of cooperation studies focused on the framework of repeated games, in which dynamic environmental state transitions are not provided. Yet, from a reinforcement learning perspective, learning in multi-state environments is standard. On the other hand, many established reinforcement algorithms are highly stochastic and require considerate training time. Therefore it is generally hard to understand how and what an agent has learned.

Hence, this chapter will present a novel methodological extension, refining techniques of the statistical physics literature on learning dynamics to derive a deterministic limit of a general class of reinforcement learning algorithms, called temporal difference learning. The focus of this chapter is on the derivation and analysis of the learning equations, not their application to a social-ecological dilemma environment. This will be the topic of Chapter 5. Instead, the method's potential is demonstrated with the three well established learning algorithms Q learning, SARSA learning and Actor-Critic learning. Illustrations of their dynamics on previously used multi-agent, multi-state environments reveal a wide range of different dynamical regimes, such as convergence to fixed points, limit cycles and even deterministic chaos.

This chapter is based on parts of (Barfuss et al., 2019, P7).

4.1 Introduction

Individual learning through reinforcements is a central approach in the fields of artificial intelligence (Busoniu et al., 2008; Sutton and Barto, 1998; Wiering and Otterlo, 2012), neuroscience (Hassabis et al., 2017; Shah, 2012), learning in games

(Fudenberg and Levine, 1998) and behavioral game theory (Camerer and Ho, 1999; Camerer, 2003; Erev and Roth, 1998; Roth and Erev, 1995), thereby offering a general purpose principle to either solve complex problems or explain behavior. Also in the fields of complexity economics (Arthur, 1993, 1999) and social science (Macy and Flache, 2002), reinforcement learning has been used as a model for human behavior to study social dilemmas.

Dynamical systems understanding. However, there is a need for improved understanding and better qualitative insight into the characteristic dynamics that different learning algorithms produce. Therefore, reinforcement learning has also been studied from a dynamical systems perspective. In their seminal work, Börgers and Sarin (1997) showed that one of the most basic reinforcement learning update schemes, Cross learning (Cross, 1973), converges to the replicator dynamics of evolutionary games theory in the continuous time limit. This has led to at least two, presumably non-overlapping research communities, one from statistical physics (Aloric et al., 2016; Bladon and Galla, 2011; Galla, 2009, 2011; Galla and Farmer, 2013; Marsili et al., 2000; Realpe-Gomez et al., 2012; Sanders et al., 2012; Sato and Crutchfield, 2003; Sato et al., 2002, 2005), and one from computer science machine learning (Bloembergen et al., 2015; Hennes et al., 2009, 2010; Kaisers and Tuyls, 2010; Tuyls and Nowé, 2005; Tuyls and Parsons, 2007; Tuyls et al., 2003, 2006; Vrancx et al., 2008). Thus, Sato and Crutchfield (2003) and Tuyls et al. (2003) independently deduced identical learning equations in the year 2003.

The statistical physics community usually considers the deterministic limit of the stochastic learning equations, assuming infinitely many interactions between the agents before an adaptation of behavior occurs. This limit can either be performed in continuous time with differential equations (Sato and Crutchfield, 2003; Sato et al., 2002, 2005) or discrete time with difference equations (Bladon and Galla, 2011; Galla, 2009, 2011). The differences between both variants can be significant (Galla, 2011; Realpe-Gomez et al., 2012). Deterministic chaos was found to emerge when learning simple (Sato et al., 2002) as well as complicated games (Galla and Farmer, 2013). Relaxing the assumption of infinitely many interactions between behavior updates revealed that noise can change the attractor of the learning dynamics significantly, e.g. by noise-induced oscillations (Galla, 2009, 2011).

However, these statistical physics studies so far considered only repeated normal form games. These are games where the payoff depends solely on the set of current actions, typically encoded in the entries of a payoff matrix (for the typical case of two players). Receiving payoff and choosing another set of joint actions is performed repeatedly. This setup lacks the possibility to study dynamically changing environments and their interplay with multiple agents. In those systems, rewards do not depend only on the joint action of agents, but also on the states of the environment. Environmental state changes may occur probabilistically and depend also on joint actions and the current state. Such a setting is also known as a Markov game

or stochastic game (Neyman and Sorin, 2003; Shapley, 1953). Thus, a repeated normal form game is a special case of a stochastic game with only one environmental state. Notably Akiyama and Kaneko (2000, 2002) did emphasize the importance of a dynamically changing environment, however did not utilize a reinforcement learning update scheme.

The computer science machine learning community dealing with reinforcement learning as a dynamical system (see Bloembergen et al. (2015) for an overview) particularly emphasizes the link between evolutionary game theory and multi-agent reinforcement learning as a well grounded theoretical framework for the latter (Bloembergen et al., 2015; Tuyls and Nowé, 2005; Tuyls and Parsons, 2007; Tuyls et al., 2006). This dynamical systems perspective is proposed as a way to gain qualitative insights about the variety of multi-agent reinforcement learning algorithms (see Busoniu et al. (2008) for a review). Consequently, this literature developed a focus on the translation of established reinforcement learning algorithms to a dynamical systems description, as well as the development of new algorithms based on insights of a dynamical systems perspective. While there is more work on stateless games (e.g. Q learning (Tuyls et al., 2003), frequency adjusted multi-agent Q learning (Kaisers and Tuyls, 2010)), multi-agent learning dynamics for multi-state environments have been developed as well, such as the piecewise replicator dynamics (Vrancx et al., 2008), the state-coupled replicator dynamics (Hennes et al., 2009) or the reverse engineering state-coupled replicator dynamics (Hennes et al., 2010).

Research challenge. Both communities, statistical physics and machine learning, share the interest in better qualitative insight into multi-agent learning dynamics. While the statistical physics community focuses more on dynamical properties the same set of learning equations can produce, it leaves a research gap of learning equations capable of handling multiple environmental states. The machine learning community on the other hand aims more towards algorithm development, but so far put their focus less on a dynamical systems understanding. Taken together, there is the challenge of developing a dynamical systems theory of multi-agent learning dynamics in varying environmental states.

Overview. This chapter aims to contribute to such a dynamical systems theory of multi-agent learning dynamics. It presents a novel methodological extension for obtaining the deterministic limit of multi-state temporal difference reinforcement learning, based on the interaction-adaptation timescales separation. In essence, it consists of formulating the temporal difference error for batch learning, and sending the batch size to infinity. This method is performed on the three prominent learning algorithms of Q learning, SARSA learning and Actor-Critic learning. Illustrations of their learning dynamics reveal multiple different dynamical regimes, such as fixed points, periodic orbits and deterministic chaos.

In Sec. 4.2 temporal difference reinforcement learning is introduced, based on the foundations presented in Chapter 3. Sec. 4.3 then presents the novel methodological extension to obtain the deterministic limit of temporal difference reinforcement learning, and demonstrates it for multi-state Q learning, SARSA learning and Actor-Critic learning. Sec. 4.4 illustrates their learning dynamics with previously used two-agents two-actions two-states environments, before Sec. 4.5 summarizes the main findings of this chapter.

4.2 Background: Temporal difference reinforcement learning

In contrast to the typical game theoretic assumption of perfect information agents are assumed to know nothing about the game in advance. They can only gain information about the environment and other agents through interactions. They do not know the true reward tensor $\mathbf{R}$ or the true transition probabilities $T_{s\mathbf{a}s'}$ (see Chapter 3 for the mathematical foundations). They experience only reinforcements (i.e. particular rewards $R^i_{s\mathbf{a}s'}$), while observing the current true Markov state of the environment.

In essence, reinforcement learning consists of iterative behavior changes towards a behavior profile with maximum state-values. However, due to the agents' limited information about the environment, they generally cannot compute a behavior profile's true state- and state-action values, $V^i_s(\mathbf{X})$ (Eq. 3.17) and $Q^i_{sa}(\mathbf{X})$ (Eq. 3.19), as defined in the previous section. Therefore, agents use time dependent *state-* and *state-action-value approximations*, $\tilde{V}^i_s(t)$ and $\tilde{Q}^i_{sa}(t)$, during the reinforcement learning process.

4.2.1 Temporal difference error

Basically, state-action-value approximations $\tilde{Q}^i_{sa}$ get iteratively updated by a temporal difference error $TD^i_{sa}(t)$:

$$\tilde{Q}^i_{sa}(t+1) = \tilde{Q}^i_{sa}(t) + \alpha^i TD^i_{sa}(t), \tag{4.1}$$

with $\alpha^i \in (0,1)$ being the *learning rate* of agent i.

The temporal difference error expresses a difference in the estimation of state-action values. New experience is used to compute a new estimate of the current state-action-value and corrected by the old estimate. The estimate from the new experience uses exactly the recursive relation of value functions from the Bellmann equation (Eq. 3.14),

$$TD^i_{sa}(t) = \delta_{ss(t)}\delta_{aa(t)} \cdot \Bigg[\underbrace{(1-\gamma^i) R^i_{s(t)a(t)\mathbf{a}^{-i}(t)s(t+1)} + \gamma^i \, \Upsilon^i_{s(t+1)}\,(t)}_{\text{estimate from new experience}} - \underbrace{\Upsilon^i_{s(t)}(t)}_{\text{old estimate}} \Bigg]. \tag{4.2}$$

Here, s and a denote the state-action pair whose temporal difference error is calculated. With $s(t)$, $a(t)$, etc. the state, action, etc. that occurred at time step t is referred to. Thus, the notation $R^i_{s(t)a(t)\mathbf{a}^{-i}(t)s(t+1)}$ denotes the entry of the reward tensor $R^i_{sa\mathbf{a}^{-i}s'}$ when at time step t the environmental state was s ($s(t) = s$), agent i chose action a ($a(t) = a$), the other agents chose the joint action $\mathbf{a}^{-i}$ ($\mathbf{a}^{-i}(t) = \mathbf{a}^{-i}$) and the next environmental state was s' ($s(t+1) = s'$). The $\Upsilon^i_{s(t+1)}(t)$ indicates the state-value estimate at time step t of the state visited at the next time step $s(t+1)$. $\Upsilon^i_{s(t)}(t)$ denotes the value estimate at time step t of the current state $s(t)$. Different choice for these value estimations are possible, leading to different learning variants (see below).

The Kronecker deltas $\delta_{ss(t)}, \delta_{aa(t)}$ indicate that the temporal difference error for state-action pair (s, a) is only non-zero when (s, a) was actually visited in time step t. This denotes and emphasizes that agents can only learn from experience. In contrast, e.g. experience-weighted-attraction learning (Camerer and Ho, 1999) assumes that action-value approximations can be updated with hypothetical rewards an agent would have received if it had played a different than the current action. These two cases have been referred to as *full vs. partial information* (Marsili et al., 2000). Thus, the Kronecker deltas in Eq. 4.2 indicate a partial information update. The agents use only information experienced through interaction.

The state-action-value approximations $\tilde{Q}^i_{sa}$ are translated to a behavior profile according to the Gibbs-Boltzmann distribution (Sutton and Barto, 1998)

$$X^i_{sa}(t) = \frac{\exp(\beta^i \tilde{Q}^i_{sa}(t))}{\sum_b \exp(\beta^i \tilde{Q}^i_{sb}(t))}, \tag{4.3}$$

which is also called softmax. The behavior profile $\mathbf{X}$ becomes a dynamic variable as well. The parameter β^i controls the *intensity of choice* or the *exploitation level* of agent i controlling the *exploration-exploitation trade-off*. In analogy to statistical physics, β^i is the inverse temperature. For high β^i agents tend to exploit their learned knowledge about the environment, leaning towards actions with high estimated state-action value. For low β^i agents are more likely to deviate from these high value

actions in order to explore the environment further with the chance of finding actions, which eventually might lead to even higher values. Other behavior profile translations exist as well, e.g. ϵ-greedy (Sutton and Barto, 1998).

4.2.2 Three learning variants

The specific choices of the value estimates Υ in the temporal difference error result in different reinforcement learning variants.

Q learning. For the Q learning algorithm (Sutton and Barto, 1998; Wiering and Otterlo, 2012) $\Upsilon^i_{s(t+1)}(t) = \max_b \tilde{Q}^i_{s(t+1)b}(t)$ and $\Upsilon^i_{s(t)}(t) = \tilde{Q}^i_{s(t)a(t)}(t)$. Thus, the Q learning update takes the maximum of the next state-action-value approximations as an estimate for the next state-value, regardless of the actual next action the agent plays. This is reasonable because the maximum is the highest value achievable given the current knowledge of the agent. For the state-value estimate of the current state, the Q learner takes the current state-action-value approximation $Q^i_{s(t)a(t)}(t)$. This is reasonable because this is exactly the quantity that gets updated by Eq. 4.1.

SARSA learning. For SARSA learning (Sutton and Barto, 1998; Wiering and Otterlo, 2012) $\Upsilon^i_{s(t+1)}(t) = \tilde{Q}^i_{s(t+1)a(t+1)}(t)$ and $\Upsilon^i_{s(t)}(t) = \tilde{Q}^i_{s(t)a(t)}(t)$, where $a(t+1)$ denotes the action taken by agent i at the next time step. Thus, the SARSA algorithm uses the five ingredients of an update sequence of State, Action, Reward, next State, next Action to perform one update. In practice, the SARSA sequence has to be shifted one time step back to know what the actual "next" action of the agent was.

Actor-Critic (AC) learning. For AC learning (Sutton and Barto, 1998; Wiering and Otterlo, 2012) $\Upsilon^i_{s(t+1)}(t) = \tilde{V}^i_{s(t+1)}(t)$ and $\Upsilon^i_{s(t)}(t) = \tilde{V}^i_{s(t)}(t)$. Compared to Q and SARSA learning, it has an additional data structure of state-value approximations, which get separately updated according to $\tilde{V}^i_s(t+1) = \tilde{V}^i_s(t) + \alpha^i \cdot TD^i_{sa}(t)$. The state-action-value approximations $\tilde{Q}^i_{sa}$ serve as the actor which get criticized by the state-value approximations $\tilde{V}^i_s$.

Tab. 4.1a summarizes the values estimates Υ for these three learning variants. Q and SARSA learning are structurally more similar compared to the Actor-Critic learner, which uses an additional data structure of state-value approximations $\tilde{V}^i_s$.

4.3 Deterministic limit

The previous section gave only a brief introduction to temporal difference reinforcement learning. A more comprehensive presentation can be found in the excellent book by Sutton and Barto (1998).

	$\Upsilon^i_{s(t+1)}(t)$	$\Upsilon^i_{s(t)}(t)$
Q	$\max_b \tilde{Q}^i_{s(t+1)b}(t)$	$\tilde{Q}^i_{s(t)a(t)}(t)$
SARSA	$\tilde{Q}^i_{s(t+1)a(t+1)}(t)$	$\tilde{Q}^i_{s(t)a(t)}(t)$
AC	$\tilde{V}^i_{s(t+1)}(t)$	$\tilde{V}^i_{s(t)}(t)$

(a) K=1

	$\Upsilon^i_{s(t+1)}(t)$	$\Upsilon^i_{s(t)}(t)$
Q	${}^{\max}\mathcal{Q}^i_{sa}(\mathbf{X})$	$\frac{1}{\beta^i}\log X^i_{sa}(t)$
SARSA	${}^{\text{next}}\mathcal{V}^i_{sa}(\mathbf{X})$	$\frac{1}{\beta^i}\log X^i_{sa}(t)$
AC	${}^{\text{next}}\mathcal{V}^i_{sa}(\mathbf{X})$	/

(b) $K = \infty$

Table 4.1: Overview of the three reinforcement learning variants: Q learning, SARSA learning and Actor-Critic (AC) learning. Shown in the columns are the value estimates for the next state $\Upsilon^i_{s(t+1)}(t)$ and the current state $\Upsilon^i_{s(t)}(t)$ for both ends of the batch size spectrum: $K = 1$ and $K = \infty$.

This section will present a novel extension to the methodology of interaction-adaptation timescales separation to the general class of temporal difference reinforcement learning. In summary, i) a batch formulation of the temporal difference error will be given, ii) the batch size is sent to infinity, separating the timescales of interaction and adaptation and iii) a resulting deterministic limit conversion rule for discrete time updates is presented.

This method is then showcased with the three learning variants of Q, SARSA and Actor-Critic learning. For the statistical physics community, the novelty consists of learning equations, capable of handling environmental state transitions. For the machine learning community the novelty lies in the systematic methodology used to obtain the deterministic learning equations. Note that these deterministic learning equations will not depend on the state- or state-action-value approximations anymore, being iterated maps of the behavior profile alone.

Following e.g. Bladon and Galla (2011), Sato and Crutchfield (2003), and Sato et al. (2005), Eqs. 4.1 and 4.3 are combined to

$$X^i_{sa}(t+1) = \frac{X^i_{sa}(t)\exp\left(\alpha^i\beta^i TD^i_{sa}(t)\right)}{\sum_b X^i_{sb}(t)\exp\left(\alpha^i\beta^i TD^i_{sb}(t)\right)}. \tag{4.4}$$

Although it appears that only the product $\alpha^i\beta^i$ matters for a behavior profile update, the temporal difference error TD^i_{sa} may depend only on the exploitation level β^i, as shown below.

Next, the temporal difference error for batch learning is presented.

4.3.1 Batch learning

Batch learning means that several time steps of interaction with the environment and the other agents take place before an update of the state-action-value approximations

and the behavior profile occurs. It has also been interpreted as a form of history replay (Lange et al., 2012) which is essential to stabilize the learning process when function approximation (e.g. by deep neural networks) is used (Mnih et al., 2015). History (i.e. already experienced state, action, next state triples) is used again for an update of the state-action-value approximations.

Imagine that the information from these interactions are stored inside a batch of size $K \in \mathbb{N}$. The corresponding temporal difference error of batch size K reads:

$$\begin{aligned} TD^i_{sa}(t;K) :=& \frac{1}{K(s,a)} \sum_{k=0}^{K-1} \Big[\delta_{ss(t+k)} \delta_{aa(t+k)} \\ & \cdot \Big((1-\gamma^i) R^i_{s(t+k)a(t+k)\mathbf{a}^{-i}(t+k)s(t+k+1)} \\ & + \gamma^i \, \Upsilon^i_{s(t+k+1)}(t) - \Upsilon^i_{s(t)}(t) \Big) \Big] \end{aligned} \tag{4.5}$$

where $K(s,a) = \max(1, \sum_{k=0}^{K-1} \delta_{ss(t+k)} \delta_{aa(t+k)})$ denotes the number of times the state-action pair (s,a) was visited. If the state-action pair (s,a) was never visited, $K(s,a) = 1$. The agents interact K times under the same behavior profile and use the sample average to summarize the new experience in order to update the state-action-value approximations:

$$\tilde{Q}^i_{sa}(t+K) = \tilde{Q}^i_{sa}(t) + \alpha^i TD^i_{sa}(t;K). \tag{4.6}$$

The notation $TD^i_{sa}(t)$ is short for a batch update of batch size 1: $TD^i_{sa}(t) = TD^i_{sa}(t;1)$.

4.3.2 Separation of timescales

The deterministic limit of the temporal difference learning dynamics is obtained by sending the batch size to infinity, $K \to \infty$. Equivalently, this can be regarded as a separation of timescales. Two processes can be distinguished during an update of the state-action-value approximations $\Delta\tilde{Q}^i_{sa}(t) := \tilde{Q}^i_{sa}(t+1) - \tilde{Q}^i_{sa}(t)$: adaptation and interaction,

$$\begin{aligned} \Delta\tilde{Q}^i_{sa}(t) = & \, \alpha^i \delta_{ss(t)} \delta_{aa(t)} \\ & \cdot \overbrace{\Big(\underbrace{(1-\gamma^i) R^i_{s(t)a(t)\mathbf{a}^{-i}(t)s(t+1)} + \gamma^i \, \Upsilon^i_{s(t+1)}(t)}_{\text{interaction}} - \Upsilon^i_{s(t)}(t) \Big)}^{\text{adaptation}}. \end{aligned} \tag{4.7}$$

By separating the timescales of both processes, one assumes that (infinitely) many interactions happen before one step of behavior profile adaptation occurs.

Under this assumption one can replace the sample average, i.e. the sum over sequences of states and actions with the behavior profile average, i.e. the sum over state-action behavior and transition probabilities according to

$$\boxed{\frac{1}{K(s,a)}\sum_{k=0}^{K-1}\delta_{ss(t+k)}\delta_{aa(t+k)} \to \sum_{s'}\sum_{\mathbf{a}^{-i}}\mathbf{X}^{-i}_{s\mathbf{a}^{-i}}T_{sa\mathbf{a}^{-i}s'}.} \tag{4.8}$$

This is only true under the assumption of an ergodic transition tensors (c.f. Chapter 3). For example, the immediate reward $R^i_{s(t)a(t)\mathbf{a}^{-\mathbf{i}}(t)s(t+1)}$ in the temporal difference error becomes ${}_{\mathbf{TX}^{-i}}\langle R\rangle^i_{sa}$. The time t gets rescaled accordingly as well. Taking the limit $K \to \infty$ in this way, the learning equations remain in discrete time, leaving the continuous time limit following e.g. Galla and Farmer (2013), Sato and Crutchfield (2003), and Sato et al. (2005) for future work.

4.3.3 Three learning variants

Next, the deterministic limits of the temporal difference error of the three learning variants of Q, SARSA and Actor-Critic learning are presented. Inserting them into Eq. 4.4 yields the complete description of the behavior profile update in the deterministic limit. Tab. 4.1 presents an overview of the resulting equations and a comparison to their batch size $K = 1$ versions.

Q learning

The temporal difference error of Q learning consists of three terms: i) $R^i_{s(t)a(t)\mathbf{a}^{-\mathbf{i}}(t)s(t+1)}$ ii) $\max_b \tilde{Q}^i_{s(t+1)b}(t)$ and iii) $\tilde{Q}^i_{s(t)a(t)}(t)$. As already stated $R^i_{s(t)a(t)\mathbf{a}^{-\mathbf{i}}(t)s(t+1)} \to {}_{\mathbf{TX}^{-i}}\langle R\rangle^i_{sa}$ under $K \to \infty$. $\max_b \tilde{Q}^i_{s(t+1)b}(t) \to {}^{\max}\mathcal{Q}^i_{sa}(\mathbf{X})$ which is defined as

$${}^{\max}\mathcal{Q}^i_{sa}(\mathbf{X}) := \sum_{s'}\sum_{\mathbf{a}^{-i}}\mathbf{X}^{-i}_{s\mathbf{a}^{-i}}T_{sa\mathbf{a}^{-i}s'}\max_b Q^i_{s'b}(\mathbf{X}) \tag{4.9}$$

using the deterministic limit conversion rule (Eq. 4.8). Because of the assumption of infinite interactions, one can here replace the state-action-value approximations $\tilde{Q}^i_{s(t+1)b}$ by the true state-action-values $Q^i_{s'b}$ as defined by Eq. 3.19. Note that in Eq. 4.9 this true state-action-value carries the state-index s' returning the maximum state-action value of the *next* state.

For the third term, Eq. 4.3 is inverted, yielding $\tilde{Q}^i_{sa}(t) = (\beta^i)^{-1}\log X^i_{sa}(t) + \text{const}^i_s$, where const^i_s is constant in actions, but may vary for each agent and state. Now, one can easily show that the dynamics induced by Eq. 4.4 are invariant against additive transformations in the temporal difference error $TD^i_{sa}(t,\infty) \to TD^i_{sa}(t,\infty) + \text{const}^i_s$. Thus, the third term can be converted according to $\tilde{Q}^i_{s(t)a(t)}(t) \to (\beta^i)^{-1}\log X^i_{sa}(t)$.

All together, the temporal difference error for Q learning in the deterministic limit reads

$$ {}^{\mathrm{q}}TD^i_{sa}(t,\infty) = (1-\gamma^i)\,{}_{\mathbf{TX}^{-i}}\langle R\rangle^i_{sa} + \gamma^i \cdot {}^{\max}\mathcal{Q}^i_{sa}(\mathbf{X}) - \frac{1}{\beta^i}\log X^i_{sa}(t). \tag{4.10} $$

SARSA learning

Two of the three terms of the SARSA temporal difference error are identical to the one of Q learning, leaving $\tilde{Q}^i_{s(t+1)a(t+1)}(t)$ which is replaced by

$$ {}^{\mathrm{next}}\mathcal{Q}^i_{sa}(\mathbf{X}) := \sum_{s'}\sum_{\mathbf{a}^{-i}} \mathbf{X}^{-i}_{s\mathbf{a}^{-i}} T_{sa\mathbf{a}^{-i}s'} \sum_b X^i_{s'b} Q^i_{s'b}(\mathbf{X}) \tag{4.11} $$

using again the deterministic limit conversion rule (Eq. 4.8) and the state-action-value $Q^i_{s'b}(\mathbf{X})$ of the behavior profile $\mathbf{X}$ according to Eq. 3.19. Thus, the temporal difference error for the SARSA learning update in the deterministic limit reads

$$ {}^{\mathrm{sarsa}}TD^i_{sa}(t;\infty) = (1-\gamma^i)\,{}_{\mathbf{TX}^{-i}}\langle R\rangle^i_{sa} + \gamma^i \cdot {}^{\mathrm{next}}\mathcal{Q}^i_{sa}(\mathbf{X}) - \frac{1}{\beta^i}\log X^i_{sa}(t). \tag{4.12} $$

Actor-Critic (AC) learning

For the temporal difference error for AC learning one has to find replacements for i) $\tilde{V}^i_{s(t+1)}(t)$ and ii) $\tilde{V}^i_{s(t)}(t)$. Applying again Eq. 4.8 yields $\tilde{V}^i_{s(t+1)}(t) \to {}^{\mathrm{next}}\mathcal{V}^i_{sa}$ defined as

$$ {}^{\mathrm{next}}\mathcal{V}^i_{sa} = \sum_{s'}\sum_{\mathbf{a}^{-i}} \mathbf{X}^{-i}_{s\mathbf{a}^{-i}} T_{sa\mathbf{a}^{-i}s'} V^i_{s'}(\mathbf{X}), \tag{4.13} $$

using Eq. 3.17 for the state-value $V^i_{s'}(\mathbf{X})$. Eq. 4.13 gives the average value of the next state given that in the current state the agent took action a. From Eq. 3.20 immediately follows ${}^{\mathrm{next}}\mathcal{V}^i_{sa}(\mathbf{X}) = {}^{\mathrm{next}}\mathcal{Q}^i_{sa}(\mathbf{X})$ from the SARSA update.

The second remaining term belongs to the slower adaptation timescale, or in other words: occurs outside the batch. Thus, the deterministic limit conversion rule (Eq. 4.8) does not apply. One could think of a conversion $\tilde{V}^i_{s(t)}(t) := \sum_a X^i_{sa}\tilde{Q}^i_{s(t)a(t)}(t) \to (\beta^i)^{-1}\sum_a X^i_{sa}(t)\log X^i_{sa}(t)$. However, the remaining term is constant in action, and therefore irrelevant for the dynamics, as argued above. Thus, one can simply put $\tilde{V}^i_{s(t)}(t) \to 0$.

All together, the temporal difference error of the Actor-Critic learner in the deterministic limit reads

$$ {}^{\mathrm{ac}}TD^i_{sa}(t,\infty) = (1-\gamma^i)\,{}_{\mathbf{TX}^{-i}}\langle R\rangle^i_{sa} + \gamma^i\,{}^{\mathrm{next}}\mathcal{V}^i_{sa}(\mathbf{X}) \tag{4.14} $$

4.3.4 Derivation of the Jacobian

The following sections present the derivation of the Jacobian matrix. This is done in order to compute Lyapunov exponents using an iterative QR decomposition according to Sandri (1996). Readers not interested in these mathematical details are invited to safely skip this section.

Eq. 4.4 constitutes a map f, which iteratively updates the behavior profile $\mathbf{X} \in \mathbb{R}^{N\times M\times Z}$. Consequently, one can represent its derivative as a Jacobian tensor $f'(\mathbf{X}) \in \mathbb{R}^{N\times M\times Z\times N\times M\times Z}$.

Let $A^i_{sa} := X^i_{sa} \exp\left(\alpha^i \beta^i T D^i_{sa}(\mathbf{X})\right)$ be the numerator of Eq. 4.4, and $B^i_s := \sum_b A^i_{sb}$ its denominator, i.e. $f =: A/B$. Hence,

$$f'(\mathbf{X}) = \frac{A'B - B'A}{B^2} \tag{4.15}$$

or, more precisely, in components,

$$\frac{df^i_{sa}(\mathbf{X})}{dX^j_{rb}} = \frac{\frac{dA^i_{sa}(\mathbf{X})}{dX^j_{rb}} B^i_s(\mathbf{X}) - \frac{dB^i_s}{dX^j_{rb}}(\mathbf{X}) A^i_{sa}(\mathbf{X})}{(B^i_s(\mathbf{X}))^2}. \tag{4.16}$$

A and B are known, and if A' is known, B' is easily obtained by $\frac{dB^i_s(\mathbf{X})}{dX^j_{rb}} = \sum_c \frac{dA^i_{sc}(\mathbf{X})}{dX^j_{rb}}$. Therefore one needs to compute A' for the three types: Q, SARSA and Actor-Critic learning.

Q learning

One can rewrite A^i_{sa} for the Q learner as

$$A^i_{sa} := (X^i_{sa})^{(1-\alpha^i)} \exp\left(\alpha^i \beta^i \bar{TD}^i_{sa}(\mathbf{X})\right), \tag{4.17}$$

where the estimate of the current value was removed from the temporal difference error, leaving the truncated TD error as

$$\bar{TD}^i_{sa}(\mathbf{X}) := (1-\gamma^i)\, {}_{\mathbf{TX}^{-i}}\langle R\rangle^i_{sa} + \gamma^i \max \mathcal{Q}^i_{sa}(\mathbf{X}). \tag{4.18}$$

Hence, one can write the derivative of A as

$$\begin{aligned}\frac{dA^i_{sa}(\mathbf{X})}{dX^j_{rb}} =& \exp\left(\alpha^i \beta^i \bar{TD}^i_{sa}(\mathbf{X})\right) \\ &\cdot \left((1-\alpha^i)(X^i_{sa})^{-\alpha^i} \frac{dX^i_{sa}}{dX^j_{rb}} + \alpha^i \beta^i (X^i_{sa})^{(1-\alpha^i)} \frac{d\bar{TD}^i_{sa}(\mathbf{X})}{dX^j_{rb}}\right).\end{aligned} \tag{4.19}$$

Since $\sum_c X^i_{sc} = 1$, dX^i_{sa}/dX^j_{rb} can be expressed as

$$\frac{dX^i_{sa}}{dX^j_{rb}} = \delta_{ij}\delta_{sr}(2\delta_{ab} - 1). \tag{4.20}$$

The derivative of the truncated temporal difference error reads

$$\frac{d\bar{TD}^i_{sa}(\mathbf{X})}{dX^j_{rb}} = (1-\gamma^i)\frac{d\,{}_{\mathbf{TX}^{-i}}\langle R\rangle^i_{sa}}{dX^j_{rb}} + \gamma^i \frac{d\,{}^{\max}\mathcal{Q}^i_{sa}(\mathbf{X})}{dX^j_{rb}}. \tag{4.21}$$

One can write the derivative of the reward as

$$\frac{d\,{}_{\mathbf{TX}^{-i}}\langle R\rangle^i_{sa}}{dX^j_{rb}} = \sum_{s'}\sum_{\mathbf{a}^{-i}} \frac{d\mathbf{X}^{-i}_{s\mathbf{a}^{-i}}}{dX^j_{rb}} T_{sa\mathbf{a}^{-i}s'} R^i_{sa\mathbf{a}^{-i}s'} \tag{4.22}$$

using Eq. 3.5 and Eq. 3.6, where the derivatives $d\mathbf{X}^{-i}_{s\mathbf{a}^{-i}}/dX^j_{rb}$ need to be executed according to Eq. 4.20.

For the derivative of the maximum next value one can write accordingly

$$\begin{aligned}\frac{d\,{}^{\max}\mathcal{Q}^i_{sa}(\mathbf{X})}{dX^j_{rb}} =& \sum_{s'}\sum_{\mathbf{a}^{-i}} \frac{d\mathbf{X}^{-i}_{s\mathbf{a}^{-i}}}{dX^j_{rb}} T_{sa\mathbf{a}^{-i}s'} \max_c Q^i_{s'c}(\mathbf{X}) \\ &+ \sum_{s'}\sum_{\mathbf{a}^{-i}} \mathbf{X}^{-i}_{s\mathbf{a}^{-i}} T_{sa\mathbf{a}^{-i}s'} \frac{d\max_c Q^i_{s'c}(\mathbf{X})}{dX^j_{rb}}.\end{aligned} \tag{4.23}$$

Let $a^m := \arg\max_a Q^i_{sa}(\mathbf{X})$, then

$$\frac{d\max_c Q^i_{sc}(\mathbf{X})}{dX^j_{rb}} = \delta_{aa^m}\frac{dQ^i_{sa}(\mathbf{X})}{dX^j_{rb}} \tag{4.24}$$

and

$$\frac{dQ^i_{sa}(\mathbf{X})}{dX^j_{rb}} = (1-\gamma^i)\frac{d\,{}_{\mathbf{TX}^{-i}}\langle R\rangle^i_{sa}}{dX^j_{rb}} + \gamma^i \sum_{s'} \frac{d\,{}_{\mathbf{X}}\langle T\rangle_{ss'}}{dX^j_{rb}} V^i_{s'}(\mathbf{X}) + {}_{\mathbf{X}}\langle T\rangle_{ss'} \frac{dV^i_{s'}(\mathbf{X})}{dX^j_{rb}}. \tag{4.25}$$

For the derivative of the effective Markov Chain transition tensor one can write

$$\frac{d\,{}_{\mathbf{X}}\langle T\rangle_{ss'}}{dX^j_{rb}} = \sum_{\mathbf{a}} \frac{d\mathbf{X}_{s\mathbf{a}}}{dX^j_{rb}} T_{s\mathbf{a}s'}, \tag{4.26}$$

using Eqs. 3.4 and where again the derivatives $d\mathbf{X}_{s\mathbf{a}}/dX^j_{rb}$ need to be executed according to Eq. 4.20.

For the derivative of the state value it is useful to write Eq. 3.17 as $V_s^i = (1-\gamma^i)\sum_{s'} M^{-1}_{ss'}\, {}_{\mathbf{TX}}\langle R\rangle^i_{s'}$ with $M := (\mathbb{1}_Z - \gamma^i\, {}_{\mathbf{X}}\langle T\rangle)$. Thus,

$$\frac{dV_s^i(\mathbf{X})}{dX^j_{rb}} = (1-\gamma^i)\sum_{s''}\frac{d(M^{-1}_{ss''})}{dX^j_{rb}}\, {}_{\mathbf{TX}}\langle R\rangle^i_{s''} + M^{-1}_{ss''}\frac{d\, {}_{\mathbf{TX}}\langle R\rangle^i_{s''}}{dX^j_{rb}}. \tag{4.27}$$

To obtain the derivative of the inverse matrix M^{-1} one can use $(M^{-1}M)' = 0 = (M^{-1})'M + M^{-1}M'$ and therefore $(M^{-1})' = -M^{-1}M'M^{-1}$. M' is obtained by

$$\frac{dM_{ss'}}{dX^j_{rb}} = -\gamma^i\frac{d\, {}_{\mathbf{X}}\langle T\rangle_{ss'}}{dX^j_{rb}}. \tag{4.28}$$

The derivative of the reward is given by

$$\frac{d\, {}_{\mathbf{TX}}\langle R\rangle^i_s}{dX^j_{rb}} = \sum_{s'}\sum_{\mathbf{a}}\frac{d\mathbf{X}_{s\mathbf{a}}}{dX^j_{rb}}T_{s\mathbf{a}s'}R^i_{s\mathbf{a}s'}, \tag{4.29}$$

using Eq. 3.4 and Eq. 3.6, and where again the derivatives $d\mathbf{X}_{s\mathbf{a}}/dX^j_{rb}$ need to be executed according to Eq. 4.20.

Finally, all terms to compute the Jacobian tensor for the Q learning dynamics in their deterministic limit are given.

SARSA learning

The computation of the Jacobian tensor for the SARSA learning update in its deterministic limit is similar, except the truncated TD error reads

$$\bar{TD}^i_{sa}(\mathbf{X}) := (1-\gamma^i)\, {}_{\mathbf{TX}^{-i}}\langle R\rangle^i_{sa} + \gamma^i\, {}^{\text{next}}\mathcal{Q}^i_{sa}(\mathbf{X}). \tag{4.30}$$

instead of Eq. 4.18. Hence, for SARSA learning

$$\frac{d\bar{TD}^i_{sa}(\mathbf{X})}{dX^j_{rb}} = (1-\gamma^i)\frac{d\, {}_{\mathbf{TX}^{-i}}\langle R\rangle^i_{sa}}{dX^j_{rb}} + \gamma^i\frac{d\, {}^{\text{next}}\mathcal{Q}^i_{sa}(\mathbf{X})}{dX^j_{rb}}, \tag{4.31}$$

and

$$\begin{aligned}\frac{d\, {}^{\text{next}}\mathcal{Q}^i_{sa}(\mathbf{X})}{dX^j_{rb}} =& \sum_{s'}\sum_{\mathbf{a}^{-i}}\frac{d\mathbf{X}^{-i}_{s\mathbf{a}^{-i}}}{dX^j_{rb}}T_{sa\mathbf{a}^{-i}s'}\sum_c X^i_{s'c}Q^i_{s'c}(\mathbf{X}) \\ &+ \sum_{s'}\sum_{\mathbf{a}^{-i}}\mathbf{X}^{-i}_{s\mathbf{a}^{-i}}T_{sa\mathbf{a}^{-i}s'}\frac{d[\sum_c X^i_{s'c}Q^i_{s'c}(\mathbf{X})]}{dX^j_{rb}}.\end{aligned} \tag{4.32}$$

The derivative of $\sum_c X^i_{s'c} Q^i_{s'c}(\mathbf{X})$ reads

$$\frac{d[\sum_c X^i_{s'c} Q^i_{s'c}(\mathbf{X})]}{dX^j_{rb}} = \sum_c \left(\frac{dX^i_{s'c}}{dX^j_{rb}} Q^i_{s'c}(\mathbf{X}) + X^i_{s'c} \frac{dQ^i_{s'c}}{dX^j_{rb}} \right). \tag{4.33}$$

All remaining terms have already been given in the previous section for the Q learning Jacobian tensor.

Actor-Critic learning

For the Actor-Critic learning update, the corresponding Eq. 4.17 reads

$$A^i_{sa} := X^i_{sa} \exp\Big(\alpha^i \beta^i TD^i_{sa}(\mathbf{X})\Big), \tag{4.34}$$

with the temporal difference error as given in Eq. 4.14. Left to obtain is the derivative of the next value estimate:

$$\frac{d\,{}^{\text{next}}\mathcal{V}^i_{sa}(\mathbf{X})}{dX^j_{rb}} = \sum_{s'} \sum_{\mathbf{a}^{-i}} \frac{d\mathbf{X}^{-i}_{s\mathbf{a}^{-i}}}{dX^j_{rb}} T_{sa\mathbf{a}^{-i}s'} V^i_{s'}(\mathbf{X}) + \sum_{s'} \sum_{\mathbf{a}^{-i}} \mathbf{X}^{-i}_{s\mathbf{a}^{-i}} T_{sa\mathbf{a}^{-i}s'} \frac{dV^i_{s'}(\mathbf{X})}{dX^j_{rb}}. \tag{4.35}$$

The derivative of the next value $V^i_{s'}$ is given by Eq. 4.27. These are all terms necessary to compute the Jacobian matrix for the Actor-Critic learning update.

4.4 Application to example environments

In the following section the derived deterministic learning equations will be applied in two different environments. Specifically, the three well established temporal difference learning variants (Q learning, SARSA learning and Actor-Critic (AC) learning) are compared in two different two-agents ($N = 2$), two-actions ($M = 2$) and two-states ($Z = 2$) environments: a two-state Matching Pennies game and a two-state Prisoner's Dilemma. Since the focus of this chapter is the derivation of the deterministic temporal difference learning equations, environments have been chosen, which have been used previously in related works (Hennes et al., 2009, 2010; Hilbe et al., 2018; Vrancx et al., 2008).

Note also that a comparison between the deterministic limit and the stochastic equations is left to future work, which presumably would add a noise term to the equations following the example of Galla (2009).

To measure the performance of an agent's behavior in a single scalar one can use the stationary state average reward $\boldsymbol{\sigma}(\mathbf{X}) \cdot {}_{\mathbf{TX}}\langle \mathbf{R} \rangle^i$, or equivalently the stationary state average value $\boldsymbol{\sigma}(\mathbf{X}) \cdot \mathbf{V}^i(\mathbf{X})$ (Eq. 3.21) as argued in Chapter 3.

In the following examples only homogeneous agents will be considered, i.e. agents whose parameters will not differ from each other. Therefore agent-indices will be dropped from α^i, β^i and γ^i for the ease of notation. Parts of Chapter 5 will explore the heterogeneous agent case.

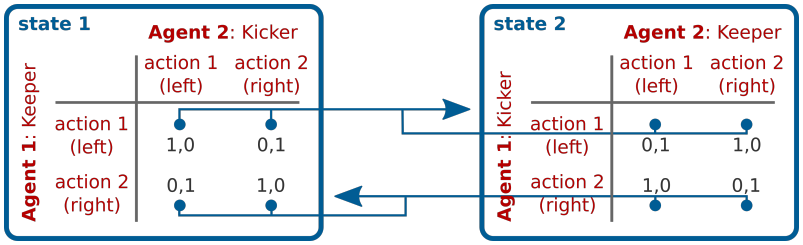

Figure 4.1: Two-state Matching Pennies environment. Rewards are given in black font in the payoff tables for each state. State transitions are indicated by blue arrows.

4.4.1 Two-state Matching Pennies

Environment description. The single state matching pennies game is a paradigmatic two-agents two-actions game. Imagine the situation of soccer penalty kicks. The keeper (agent 1) can choose to jump either to the left or right side of the goal, the kicker (agent 2) can choose to kick the ball also either to the left or the right. If both agents choose the identical side, the keeper agent wins, otherwise the kicker agent.

In the two-state version of the game according to Hennes et al. (2010) the rules are extend as follows: In state 1 the situation is as described in the single state version. Whenever agent 1 (the keeper) decides to jump to the left, the environment transitions to state 2 in which the agents switch roles: agent 1 now plays the kicker and agent 2 the keeper. From here, whenever agent 1 (now the kicker) decides to kick to the right side, the environment transitions again to state 1 and both agents switch their roles again.

Fig. 4.1 illustrate this two-state Matching Pennies games. Formally, the payoff matrices are given by

$$\begin{pmatrix} R^1_{111s'},R^2_{111s'} & R^1_{112s'},R^2_{112s'} \\ R^1_{121s'},R^2_{121s'} & R^1_{122s'},R^2_{122s'} \end{pmatrix} = \begin{pmatrix} 1{,}0 & 0{,}1 \\ 0{,}1 & 1{,}0 \end{pmatrix}$$

in state 1 and

$$\begin{pmatrix} R^1_{211s'},R^2_{211s'} & R^1_{212s'},R^2_{212s'} \\ R^1_{221s'},R^2_{221s'} & R^1_{222s'},R^2_{222s'} \end{pmatrix} = \begin{pmatrix} 0{,}1 & 1{,}0 \\ 1{,}0 & 0{,}1 \end{pmatrix}$$

in state 2 for $s' \in \{1, 2\}$. State transitions are governed by

$$\begin{pmatrix} T_{1112} & T_{1122} \\ T_{1212} & T_{1222} \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix} \text{ and } \begin{pmatrix} T_{2111} & T_{2121} \\ T_{2211} & T_{2221} \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 1 & 1 \end{pmatrix}.$$

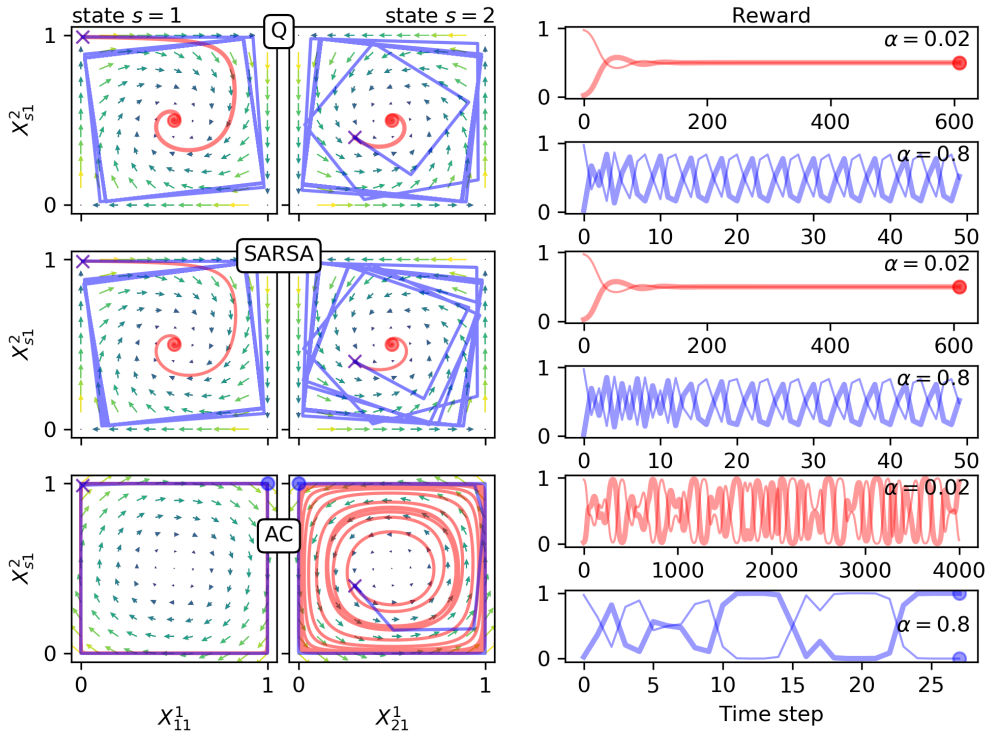

Figure 4.2: Three learners in two-state Matching Pennies environment for low farsightedness $\gamma = 0.1$; intensity of choice $\beta = 5.0$. On the left side, the temporal difference errors for the Q learner (Eq. 4.10), SARSA learner (Eq. 4.12) and Actor-Critic (AC) learner (Eq. 4.14) are shown in two behavior phase space sections, one for each state. The arrows indicate the average direction the temporal difference errors drive the learners towards, averaged over all phase space points of the other state. Arrow colors additionally encode their lengths. Selected trajectories are shown in the phase space sections, as well as by reward trajectories on the right, plotting the average reward value (Eq. 3.21) over time steps. Crosses in the phase space subsections indicate the initial behavior $(X^1_{11}, X^2_{11}, X^1_{21}, X^2_{21}) = (0.01, 0.99, 0.3, 0.4)$. Circles signal the arrival at a fixed point, determined by the absolute difference of behavior profiles between two subsequent time steps being below $\epsilon = 10^{-6}$. Trajectories are shown for two different learning rates $\alpha = 0.02$ (red) and $\alpha = 0.8$ (blue). The bold reward trajectory belongs to agent 1, the thin one to agent 2. Note that the temporal difference error is independent from the learning rate α. A variety of qualitatively different dynamical regimes can be observed.

Thus by construction, the probability of transitioning to the other state is independent of agent 2's action. Only agent 1 has agency over the state transitions. By playing a uniformly random behavior profile $(X_{11}^1, X_{11}^2, X_{21}^1, X_{21}^2) = (0.5, 0.5, 0.5, 0.5)$ both agents would obtain an average reward of 0.5 per time step.

Behavior space at low farsightedness. Fig. 4.2 compares the learners' temporal difference errors in the behavior phase space sections for each environmental state at a comparable low discount factor $\gamma \in [0, 1)$ of $\gamma = 0.1$, as well as learning trajectories for an exemplary initial condition for two learning rates $\alpha \in (0, 1)$, a low one ($\alpha = 0.02$) and a high one ($\alpha = 0.8$). Overall, one observes a variety of qualitatively different dynamical regimes, such as fixed points, periodic orbits and chaotic motion.

Specifically, one can see that Q learners and SARSA learners behave qualitatively similar in contrast to the AC learners; for both learning rates α. For the low learning rate $\alpha = 0.02$, Q and SARSA learners reach a fixed point of playing both actions with equal probability in both states, yielding a reward of 0.5. Due to the low α, this takes approx. 600 time steps. In contrast, the reward trajectory of the AC learner appears to be chaotic. Fig. 4.4 confirms this observation, which will be discussed in more detail below.

For the high learning rate $\alpha = 0.8$ both Q and SARSA learners enter a periodic limit cycle. Differences in the trajectories of Q and SARSA learners are clearly visible. The time average reward of this periodic orbit appears to be approx. 0.5 for each agent, identical to the reward of the fixed point at lower α. The AC learners, however, converge to a fixed point after oscillating near the edges of the phase space. At this fixed point in state 1 agent 1 plays action 1 with probability 1. Thus, it has trapped the system into state 2. In state 2, agent 1 plays action 2 and agent 2 plays action 1, both with full probability. Consequently agent 1 receives a reward of 1, whereas agent 2 receives zero reward. One might ask, why does agent 2 not decrease its probability for playing action 1, thereby increasing its own reward? And indeed, the arrows of the temporal difference error suggest this change of behavior profile. However, agent 2 cannot follow because its behavior is trapped on the simplex of non-zero action probabilities X_{2a}^2. For only $M = 2$ actions, $X_{21}^2 = 1$ thus cannot change anymore, regardless of the temporal difference error.

Behavior space at high farsightedness. Increasing the discount factor to $\gamma = 0.9$, one observes the learning rate α to set the timescale of learning (Fig. 4.3). The intensity of choice remained $\beta = 5.0$. A high learning rate $\alpha = 0.8$ corresponds to faster learning in contrast to a low learning rate $\alpha = 0.02$. Also the ratio of learning timescales is comparable to the inverse ratio of learning rates. For both α, Q and SARSA learners reach a fixed point, whereas the AC learners seem to move chaotically (details to be investigated below). Comparing the trajectories between the learning rates α, one observes a similar shape for each pair of learners. However, the similarity of the AC trajectories decreases at larger time steps.

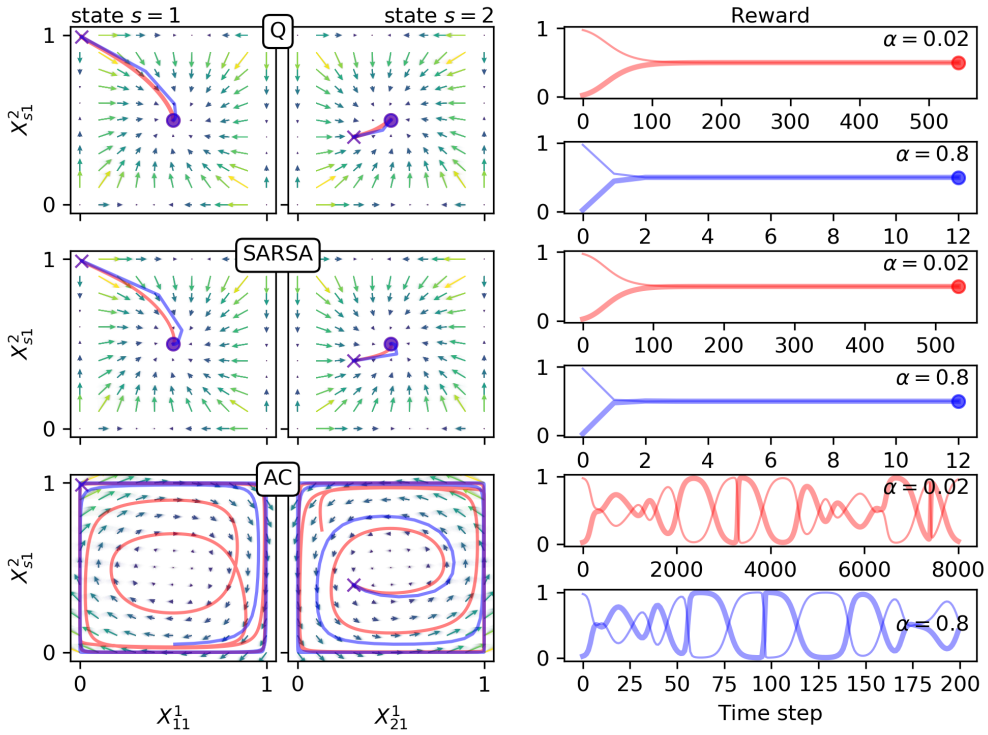

Figure 4.3: Two-state Matching Pennies environment for high farsightedness $\gamma = 0.9$; otherwise identical to Fig. 4.2.

Farsightedness and learning rate combined. So far, two parameters were varied: the farsightedness $\gamma \in [0, 1)$ and the learning rate $\alpha \in (0, 1)$. Combining Figs. 4.2 and 4.3 all four combinations of a low and a high γ with a low and a high α have been investigated. One can summarize that Q and SARSA learners converge to a fixed point for all combinations of discount factor γ and learning rate α, except when γ is low and α simultaneously high. AC dynamics seem chaotic for all combinations of α and γ.

To investigate the relationship between the parameters more thoroughly, Fig. 4.4 shows bifurcation diagrams with the bifurcation parameters α and γ. Additionally, it also gives the largest Lyapunov exponents for each learner and each parameter combination. A largest Lyapunov exponent greater than zero is a key characteristic of chaotic motion. They are computed from the analytically derived Jacobian matrix, iteratively used in a QR decomposition according to Sandri (1996).

The largest Lyapunov exponent for Q and SARSA learners align almost perfectly with each other, whereas the largest Lyapunov exponent of the AC learners behaves qualitatively different. First, consider the behavior of the Q and SARSA learners: For high learning rates α and low farsightedness γ Fig. 4.4 shows a periodic orbit with few (four) points in phase space. Largest Lyapunov exponents are distinctly below 0 at those regimes. Increasing the farsightedness γ both learners enter a regime of visiting

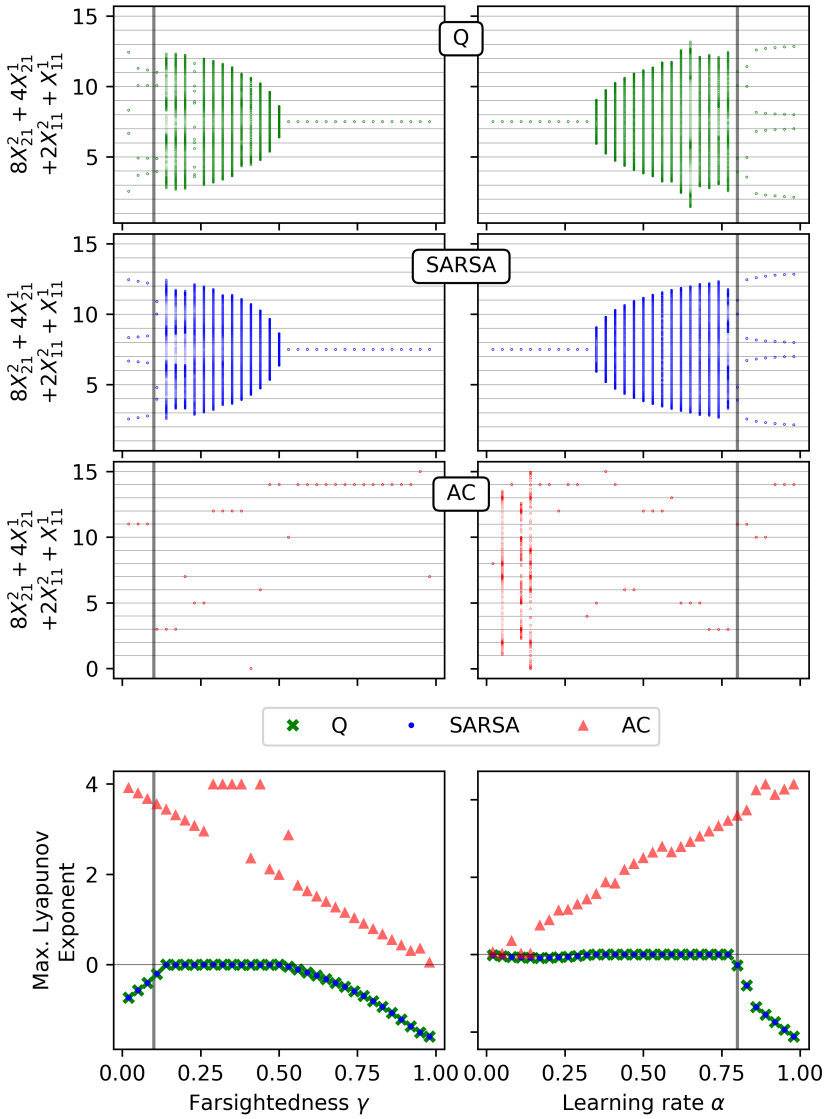

Figure 4.4: Varying farsightedness γ and learning rate α in two-state Matching Pennies environment for intensity of choice $\beta = 5.0$. On the left, the farsightedness γ is varied with learning rate $\alpha = 0.8$, as indicated by the gray vertical lines on the right. On the right, the learning rate α is varied with discount factor $\gamma = 0.1$ as indicated by the gray vertical lines on the left. The three top panels show a the visited behavior points during 1000 iterations after a transient period of 100000 time steps from initial behavior $(X_{11}^1, X_{11}^2, X_{21}^1, X_{21}^2) = (0.01, 0.99, 0.3, 0.4)$ for the Q learner (green), the SARSA learner (blue) and the Actor-Critic (AC) learner (red). Visited points are mapped to the function $8X_{21}^2 + 4X_{21}^1 + 2X_{11}^2 + X_{11}^1$ on the vertical axes to give a fuller image of the visited behavior profiles. The bottom panel shows the corresponding largest Lyapunov exponents for the three learners. Overall, Q and SARSA learner behave qualitatively more similar than the Actor-Critic learner.

many points in phase space around the stable fixed point $(X_{11}^1, X_{11}^2, X_{21}^1, X_{21}^2) = (0.5, 0.5, 0.5, 0.5)$. The largest Lyapuonv exponents are close to zero. With increasing γ the distance around this fixed point solution decreases until the dynamics converge from a farsightedness γ slightly greater than 0.5 on. From there the largest Lyapunov exponent decreases again for further increasing γ. The same observations can be made along a decreasing bifurcation parameter α, except that at the end, for low α the largest Lyapunov exponents do not decrease as distinctly as for high γ.

The behavior of the Actor-Critic dynamics is qualitatively different from the one of Q and SARSA. The placement of the fixed points on the natural numbers grid suggests that the AC learners get confined on one of the 16 (M^{NZ}) corners of the behavior phase space. No regularity to which fixed point the AC learners converge can be deduced. The largest Lyapunov exponent is always above zero and experiences an overall decreasing behavior. Similarly for a decreasing bifurcation parameter α, the largest Lyapunov exponent tends to decrease as well. Different from the bifurcation diagram along γ, for low α the system might enter a periodic motion, but only for some parameters α. No regularity can be determined at which parameters α the AC learners enter a periodic motion. A more thorough investigation of the nonlinear dynamics, especially those of the Actor-Critic learner seems of great interest, is, however, beyond the scope of this chapter and leaves promising paths for future work.

Exploitation level. Concerning the parameter β, the intensity of choice or exploitation level, one can infer from the update equations (Eq. 4.4 combined with Eq. 4.12 and Eq. 4.14), that the dynamics of the AC learners are invariant for a constant product $\alpha\beta$. This is because the temporal difference error of the Actor-Critic learner in the deterministic limit is independent of β. Further, the dynamics of the SARSA learner will converge to the dynamics of the AC learner under $\beta \to \infty$. Fig. 4.5 nicely confirms these two observations. The second observation can also be made with Tab. 4.1. Since the value estimate of the future state is identical for SARSA and AC learning, letting the value estimate of the current state vanish by sending $\beta \to \infty$ makes the SARSA learners approximate the AC learners.

As mentioned before, β controls the exploration-exploitation trade-off. In the temporal difference errors of the Q and SARSA learner it appears in the term indicating the value estimate of the current state $-1/\beta^i \log(X_{sa}^i)$. If this term dominates the temporal difference error (i.e. if β is small), the learners tend towards the center of behavior space, i.e. $(X_{11}^1, X_{11}^2, X_{21}^1, X_{21}^2) = (0.5, 0.5, 0.5, 0.5)$, forgetting what they have learned about the obtainable reward. This characteristic happens to be favorable in this two-state Matching Pennies environment, which is why Q and SARSA learners perform better in finding the $(X_{11}^1, X_{11}^2, X_{21}^1, X_{21}^2) = (0.5, 0.5, 0.5, 0.5)$ solution. On the other hand, if β is large, the temporal difference error is dominated by the current reward and future value estimate. Not being able to forget, the

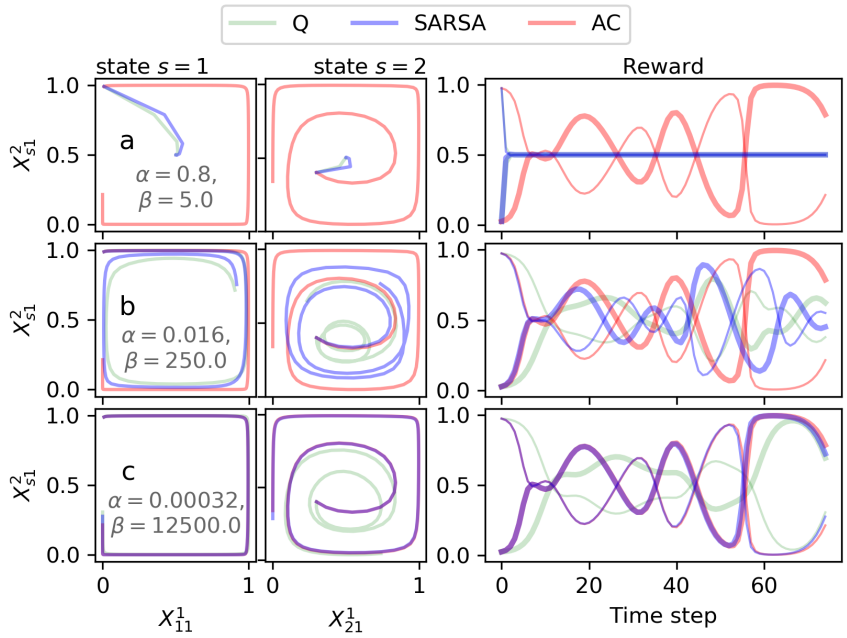

Figure 4.5: Varying exploitation level β under constant $\alpha \cdot \beta$ in two-state Matching Pennies environment for farsightedness $\gamma = 0.9$. On the left trajectories of the three learners (Q: green, SARSA: blue, Actor-Critic(AC): red) are shown in the two phase space sections, one for each state. On the right, the corresponding reward trajectories are shown. The initial behavior was $(X^1_{11}, X^2_{11}, X^1_{21}, X^2_{21}) = (0.01, 0.99, 0.3, 0.4)$. The bold reward trajectory belongs to agent 1, the thin one to agent 2. One observes the deterministic limit of Actor-Critic learning to be invariant under constant $\alpha \cdot \beta$ and SARSA learning to converge to AC learning under $\beta \to \infty$.

learners might get trapped in unfavorable behavior, as one can see observing the Actor-Critic learners. To calibrate β it is useful to make oneself clear that it must come in the unit of [log behavior] / [reward].

4.4.2 Two-state Prisoner's Dilemma

Environment description. The single state Prisoner's Dilemma is another paradigmatic two-agents, two-actions game. It has been used to model social dilemmas and study the emergence of cooperation. It describes a situation in which two prisoners are separately interrogated, leaving them with the choice to either cooperate with each other by not speaking to the police or defecting by testifying.

The two-state version, which has been used as a test environment also by Hennes et al. (2009, 2010) and Vrancx et al. (2008), extends this situation somewhat artificially by playing a Prisoner's Dilemma in each of the two states with a transition probability of 10% from one state to the other if both agents chose the same action, and a transition probability of 90% if both agents chose opposite actions.

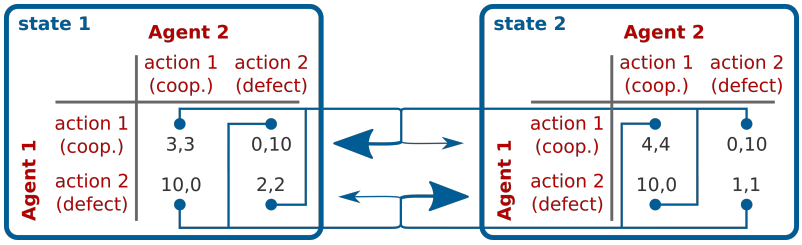

Figure 4.6: Two-state Prisoner's Dilemma environment. Rewards are given in black font in the payoff tables for each state. State transitions are indicated by blue arrows.

Fig. 4.6 illustrates these game dynamics. Formally, the payoff matrices are given by

$$\begin{pmatrix} R^1_{111s'},R^2_{111s'} & R^1_{112s'},R^2_{112s'} \\ R^1_{121s'},R^2_{121s'} & R^1_{122s'},R^2_{122s'} \end{pmatrix} = \begin{pmatrix} 3{,}3 & 0{,}10 \\ 10{,}0 & 2{,}2 \end{pmatrix}$$

in state 1 and

$$\begin{pmatrix} R^1_{211s'},R^2_{211s'} & R^1_{212s'},R^2_{212s'} \\ R^1_{221s'},R^2_{221s'} & R^1_{222s'},R^2_{222s'} \end{pmatrix} = \begin{pmatrix} 4{,}4 & 0{,}10 \\ 10{,}0 & 1{,}1 \end{pmatrix}$$

in state 2 for $s' \in \{1, 2\}$, respectively. The corresponding state transition probabilities are given by

$$\begin{pmatrix} T_{1112} & T_{1122} \\ T_{1212} & T_{1222} \end{pmatrix} = \begin{pmatrix} T_{2111} & T_{2121} \\ T_{2211} & T_{2221} \end{pmatrix} = \begin{pmatrix} 0.1 & 0.9 \\ 0.9 & 0.1 \end{pmatrix}.$$

To be precise, the rewards in each state do not resemble a classical social dilemma situation. This is because if both agents were alternately cooperating and defecting, both could receive a larger reward per time step compared to always cooperating. Hence, this stochastic game, as it was used by Hennes et al. (2009, 2010) and Vrancx et al. (2008), presents more a coordination than a cooperation challenge to the agents. The multi-state environment can here function as a coordination device. A behavior profile in which one agent exploits the other in one state, while being exploited in the other state, would result in an average reward per time step of 5 for each agent, e.g. $(X^1_{11}, X^2_{11}, X^1_{21}, X^2_{21}) = (0, 1, 1, 0)$.

Behavior space at medium farsightedness. However, for all three learning types with a medium ranged farsightedness ($\gamma = 0.45$) and an intensity of choice $\beta = 5.0$, the temporal difference error arrows are pointing on average towards the lower left defection-defection point for each state in behavior phase space (Fig. 4.7). To see whether the three learning types may converge to the described defect-cooperate, cooperate-defect solution, individual trajectories from two exemplary initial conditions and for two learning rates α are shown, as before a small one ($\alpha = 0.02$) and a high one ($\alpha = 0.8$).

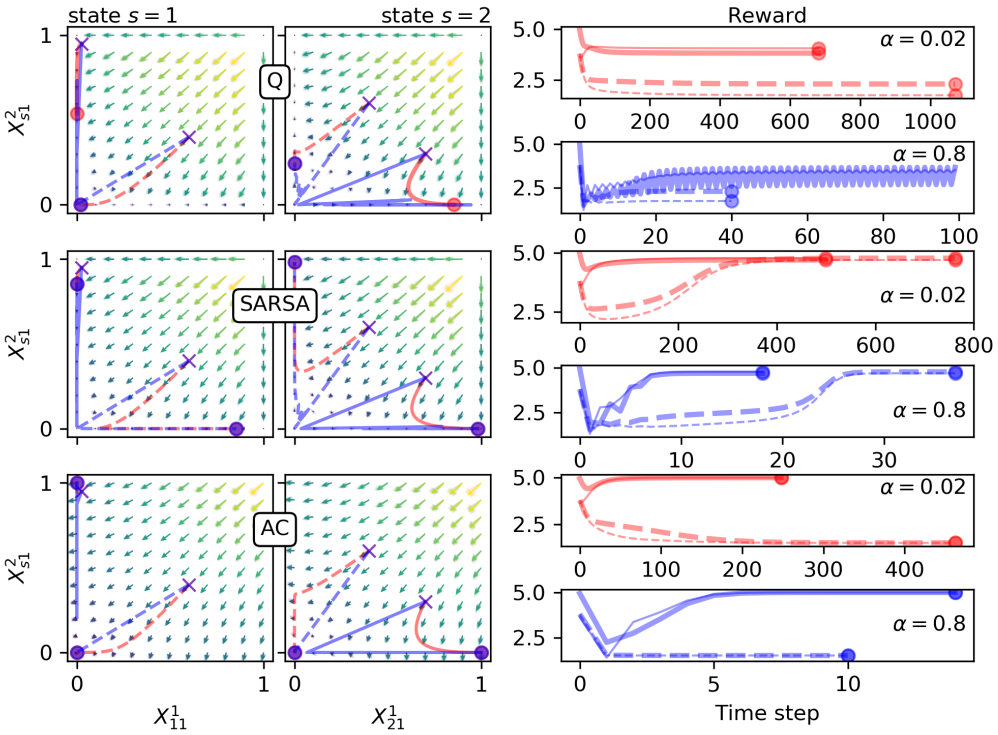

Figure 4.7: Two-state Prisoner's Dilemma environment for farsightedness $\gamma = 0.45$; otherwise identical to Fig. 4.2.

One observes qualitatively different behavior across all three learners. The Q learners converge to equilibria with average rewards distinctly below 5, the SARSA learners converge to equilibira with average rewards of almost 5 for both learning rates α and both exemplary initial conditions. Both Q and SARSA learners converge to solutions of proper probabilistic behavior, i.e. choosing action cooperate and action defect with non-vanishing chance. The Actor-Critic learners on the other hand converge to the deterministic defect-cooperate, cooperate-defect behavior described above for the initial condition shown with the non-dashed lines in Fig. 4.7 for both learning rates α (shown in red and blue). For the other exemplary initial condition, shown with the dashed lines, it converges to an all-defection solution in both states for both α.

Interestingly, for all learners, all combinations of initial conditions and learning rates converge to a fixed point solution, except for the Q learners with a comparably high learning rate $\alpha = 0.8$, which enter a periodic behavior solution for the initial condition with the non-dashed line. The same phenomenon occurred also in the Matching Pennies environment for low farsightedness $\gamma = 0.1$, however there for both, Q and SARSA learners. It seems to be caused by the comparably high learning rate. A high learning rate overshoots the behavior update resulting in a circling behavior around the fixed point. As in Fig. 4.2, the time average reward of the periodic orbit

seems to be comparable to the reward of the corresponding fixed point at lower α. Furthermore, one observes the same time re-scaling effect of the learning rate α in Fig. 4.7 as in Fig. 4.3.

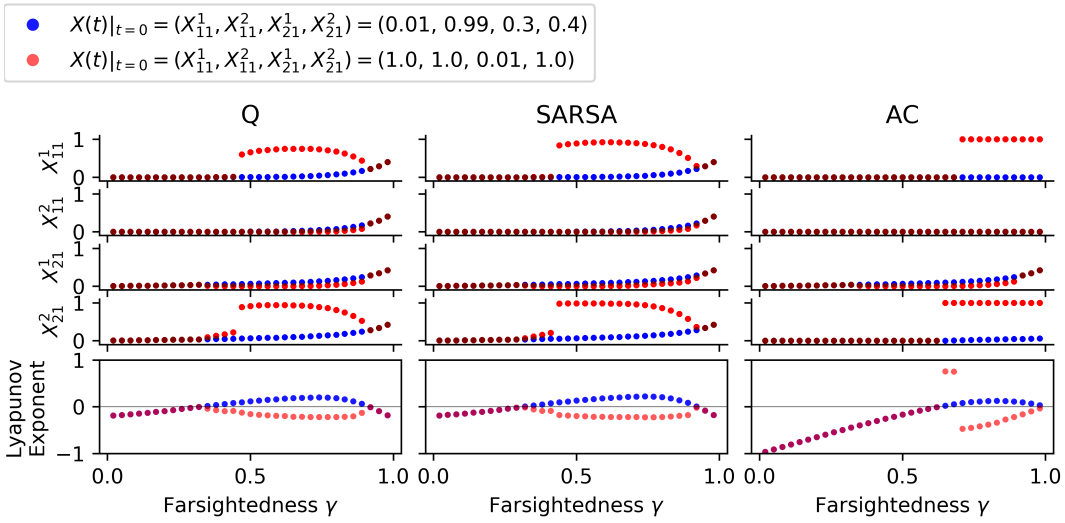

Figure 4.8: Varying farsightedness γ in two-state Prisoner's Dilemma environment for learning rate $\alpha = 0.2$ and intensity of choice $\beta = 5.0$. The four top panels show the visited behavior points $X^1_{11}, X^2_{11}, X^1_{21}, X^2_{21}$ during 1000 iterations after a transient period of 5000 time steps from initial behavior $(X^1_{11}, X^2_{11}, X^1_{21}, X^2_{21}) = (0.5, 0.5, 0.5, 0.5)$ in blue and from initial behavior $(X^1_{11}, X^2_{11}, X^1_{21}, X^2_{21}) = (0.51, 0.49, 0.49, 0.51)$ in red for the Q learner on the left, the SARSA learner in the middle and the Actor-Critic learner on the right. The bottom panel shows the corresponding largest Lyapunov exponents for the two initial conditions. Above a critical farsightedness γ all learners find the high rewarding solution from the red initial condition, but do not do so from the blue initial condition.

Varying farsightedness. To visualize the influence of the discount factor γ on the converged behavior, Fig. 4.8 shows a bifurcation diagram along the bifurcation parameter γ for two initial conditions. Dots in blue result from a uniformly random behavior profile of $(X^1_{11}, X^2_{11}, X^1_{21}, X^2_{21}) = (0.5, 0.5, 0.5, 0.5)$, whereas the dots in red initially started from the behavior profile $(X^1_{11}, X^2_{11}, X^1_{21}, X^2_{21}) = (0.51, 0.49, 0.49, 0.51)$.

Across all learners, lower discount factors γ correspond to all-defect solutions, whereas for higher γ the solutions from the initial condition shown in red tend towards the cooperate-defect, defect-cooperate solution. For low γ, the agents are less aware of the presence of other states and find the all-defect equilibrium solution of the iterated normal form Prisoner's Dilemma. The state transition probabilities have less effect on the learning dynamics. Only above a certain farsightedness, the agents find the more rewarding cooperate-defect, defect-cooperate solution.

The observation from Fig. 4.7 is confirmed that the probability to cooperate (i.e. here X^1_{11} and X^2_{21}) is lowest for the Q learners, mid range for the SARSA learners and 1 for the Actor-Critic learners. One reason for this observation can be found in

the intensity of choice parameter β. It balances the reward obtainable in the current behavior space segment with the forgetting of current knowledge to be open to new solutions. Such forgetting expresses itself by temporal difference error components pointing towards the center of behavior space. Thus, a relatively small $\beta = 5.0$ can explain why solutions at the edge of the behavior space cannot be reached by Q and SARSA learners. The AC learner misses this forgetting term in the deterministic limit and can therefore easily enter behavior profiles at the edge of the behavior space.

Q and SARSA learners have a critical discount factor γ above which the cooperate-defect, defect-cooperate high reward solution is obtained and below which the all-defect low reward solution gets selected. However, for increasing discount factors γ up to 1, Q and SARSA learners experience a drop in playing the cooperative action probability.

The Actor-Critic learners approach the cooperate-defect, defect-cooperate solution in two steps. For increasing γ, first the probability to cooperate of agent 2 in state 2 (X^2_{21}) jumps from zero to one while agent 1 still defects in state 1. Only after a slight increase of γ, agent 1 then also cooperates in state 1 (X^1_{11}).

Interestingly, for the uniformly random initial behavior condition shown in blue, there is no critical discount factor γ and no learners come close to the cooperate-defect, defect-cooperate solution. Here, only for γ close to 1, all cooperation probabilities X^i_{s1} gradually increase. Furthermore, exactly at those γ, where the cooperate-defect, defect-cooperate solution is obtained from the initial behavior condition shown in red, the solutions from the uniformly random initial behavior condition (blue) have a largest Lyapuonv exponent greater than 0. At other values of γ, the largest Lyapunov exponents for the two initial conditions overlap. This suggests that largest Lyapunov exponents greater than zero may point to the fact that other, perhaps more rewarding solutions may exist in phase space. A more thorough investigation regarding this multi-stability is an open point for future research.

Cooperation challenge. As argued above, the two-state Prisoner's Dilemma as it was used by Hennes et al. (2009, 2010) and Vrancx et al. (2008) presents rather a coordination than a cooperation challenge to the agents. Fig. 4.9 demonstrates that the derived learning dynamics are also capable to solve a cooperation challenge in a stochastic game setting, for which a two-state Prisoner's Dilemma was adapted in analogy to Hilbe et al. (2018). Fig. 4.9 confirms previous findings that cooperation emerges only in the stochastic game, compared to playing each Prisoner's Dilemma repeatedly (Hilbe et al., 2018). Further, cooperation only emerges for sufficiently large farsightedness γ.

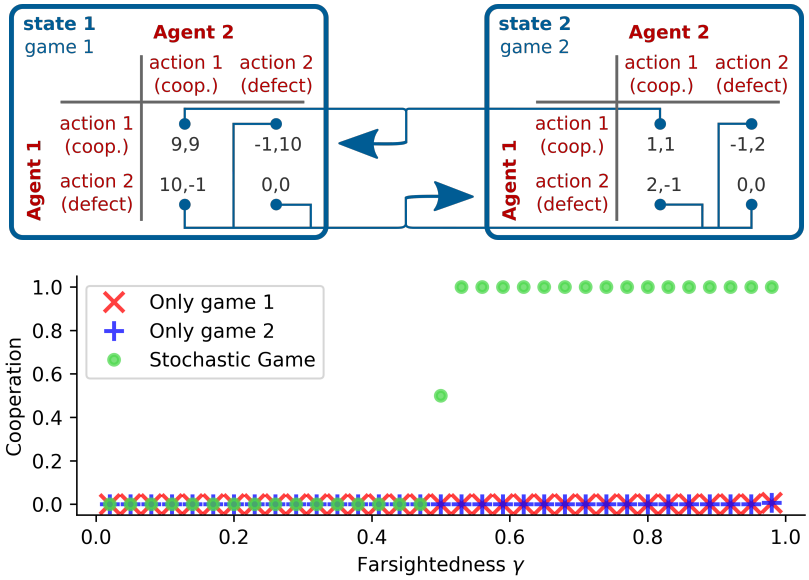

Figure 4.9: Cooperation challenge in a two-state Prisoner's Dilemma. At the top a two-state Prisoner's Dilemma game is shown, whose state games individually favor defection. At the bottom the level of cooperation is shown SARSA learners with $\alpha = 0.016$, $\beta = 250$ play after reaching a fixed point from the center of behavior space ($X^i_{sa} = 0.5$ for all i, s, a) for varying discount factors γ. Results for Q and AC learners are similar. Cooperation levels are shown for the full stochastic game as well as for each individual state game played repeatedly. For sufficiently large farsightedness cooperation can emerge in the stochastic game, in contrast to the individual repeated games.

4.5 Summary

The main contribution of this chapter is the development of a technique to obtain the deterministic limit of temporal difference reinforcement learning. Through this chapter the literature on learning dynamics from statistical physics has been combined with the evolutionary game theory-inspired learning dynamics literature from machine learning. For the statistical physics community, the novelty consists of learning equations, capable of handling environmental state transitions. For the machine learning community the novelty lies in the systematic methodology which was used to obtain the deterministic learning equations.

The presented methodology was demonstrated with the three prominent reinforcement learning algorithms from computer science: Q learning, SARSA learning and Actor-Critic learning. A comparison of their dynamics in previously used two-agents, two-actions, two-states environments has revealed the existence of a variety of qualitatively different dynamical regimes, such as convergence to fixed points, periodic orbits and deterministic chaos.

It was shown that Q and SARSA learners tend to behave qualitatively more similar in comparison to the Actor-Critic learning dynamics. This characteristic results at

least partly from the relatively low intensity of choice parameter β, controlling the exploration-exploitation trade-off via a forgetting term in the temporal difference errors. Under $\beta \to \infty$ the forgetting term vanishes and the SARSA learning dynamics approach the Actor-Critic learning dynamics.

Overall the Actor-Critic learners have a tendency to enter confining behavior profiles, due to their non-existing forgetting term. This characteristic leaves them trapped at the edges of the behavior space. In contrast, Q and SARSA learners do not show such learning behavior. Interestingly, this characteristic of the AC learners turns out to be favorable in the two-state Prisoner's Dilemma environment, where they find the most rewarding solution in more cases compared to Q and SARSA, but hinders the convergence to the fixed point solution in the two-state Matching Pennies environment. Thus, the most favorable level of forgetting depends on the environment. In order to tune the respective parameter β, the consideration that it must come in the unit of [log behavior] / [reward] may be helpful.

The effect of the learning rate α has been demonstrated, adjusting the speed of learning by controlling the amount of new information used in a behavior profile update. Thereby, within limits, α functions as a time re-scaling. However, a comparably large learning rate α might cause an overshooting phenomenon, hindering the convergence to a fixed point. Instead, the learners enter a limit cycle around that point. Nevertheless, the average reward of the limit cycling behavior was approximately equal to the one of the fixed point obtained at lower α, but took fewer time steps to reach. Thus, perhaps other dynamical regimes than fixed points, such as limit cycles or strange attractors, could be of interest in some applications of reinforcement learning.

The effect of the discount factor γ was also demonstrated, adjusting the farsightedness of the agents. At low γ the state transition probabilities have less effect on the learning dynamics compared to high discount factors.

To summarize the three parameters α, β and γ: the level of exploitation β and the farsightedness γ control *where* the learner adapts towards in behavior space, weighting current reward, expected future reward and the level of forgetting. The learning rate α controls *how fast* the learner adapts along these directions.

Outlook. This work might turn out useful for the application of reinforcement learning in various domains, with respect to parameter tuning, the design of new algorithms, and the analysis of complex strategic interactions using meta strategies, as Bloembergen et al. (2015) have pointed out. In this regard, future work could extend the presented methodology to partial observability of the Markov states of the environment (Oliehoek, 2012; Spaan, 2012), behavior profiles with history, and other-regarding agent (i.e. joint-action) learners (c.f. Busoniu et al. (2008) for an overview of other-regarding agent learning algorithms). Also, the combination of individual reinforcement learning and social learning through imitation (Bandura, 1977; Banisch and Olbrich, 2018; Barfuss et al., 2017, P1; Smolla et al., 2015) seems promising. Such endeavors would naturally lead to the exploration of network effects.

It is important to note that only a few dynamical systems reinforcement learning studies have begun to incorporate network structures between agents (Bladon and Galla, 2011; Realpe-Gomez et al., 2012).

Apart from these more technical extensions, these learning equations may turn out useful when studying the evolution of cooperation in stochastic games (Hilbe et al., 2018). With stochastic games one is able to explicitly account for a changing environment. Therefore, such studies are likely to contribute to the advancement of theoretical research on the sustainability of coupled social-ecological systems (Donges et al., 2017; Levin, 2013). Interactions, synergies and trade-offs between social (Dawes, 1980; Macy and Flache, 2002) and ecological (Heitzig et al., 2016) dilemmas can be explored using the framework of stochastic games. More realistic environments, studied with the derived learning equations, are likely to prove themselves useful in order to investigate the preconditions for sustainability: e.g. the harvesting of a common-pool renewable resource (Lindkvist and Norberg, 2014; Schill et al., 2015) or the prevention of dangerous climate change (Barrett and Dannenberg, 2012; Milinski et al., 2008). In this spirit, the next chapter will apply the derived learning equations to a particular social-ecological dilemma setting: the Ecological Public Good, modeling a social-ecological tipping element.

Python code for the reproduction of the reported results is available at github: `https://doi.org/10.5281/zenodo.1495091`.

Chapter 5

Second act: Cooperation in the ecological public good

Zwischen mir und dem Krieg ist der Erdkern, Ich sehe durch ihn hindurch. Wo noch Gras ist, sehe Ich das Gras von unten. Die Welt ist durchschaubar.

Einstein - from Paul Dessau's *Einstein*: Second act

Chapter 4 derived a deterministic limit of established so-called temporal difference reinforcement learning equations. They enable a dynamical systems perspective on the learning in multi-state environments, where the rewards of the agents depend not only on all actions but also on the current environmental state transition.

In the following chapter these derived learning dynamics will be applied to a particular environment, modeling an interlinked social-ecological dilemma. This environment, hereby called the Ecological Public Good (EcoPG), extends the established social dilemma public good by an environmental tipping element.

The Ecological Public Good within the framework of the agent-environment interface for the investigation of social-ecological systems combines two previously studied topics: i) the emergence of cooperation in stochastic games, and ii) so-called collective risk dilemmas, a particular class of games to study the cooperation for the mitigation of dangerous climate change. Due to the learning formalism presented in Chapter 4 results can be derived numerically as well as analytically. Thus, three qualitatively different parameter regimes could be identified with respect to the emergence of cooperation, depending on the parameters farsightedness, collapse risk and collapse damage.

In light of increasing political polarization this chapter asks for the stability of a cooperation agreement when one of the participating actors puts less weight on expected future gains. It can be shown that cooperation can remain stable despite considerable shortsightedness, but only if this actor's leverage to collapse the environment is large and if the expected negative impact due to the collapse is high. Conversely, this model projects that an actor who does not believe in likely and severe consequences of a tipping catastrophe will break off the cooperation agreement.

This chapter contains unpublished material. However, an independent publication based on the material presented in this chapter is planned (Barfuss et al., in prep., P8).

5.1 Introduction

Collective human action is required to steer the Earth system away from potential thresholds and stabilize it in a habitable state (Steffen et al., 2018), below planetary boundaries (Rockström et al., 2009b; Steffen et al., 2015a) and simultaneously above social foundations (Griggs et al., 2013; Raworth, 2017). In order to enter such a safe and just operating space (Raworth, 2012; Rockström et al., 2009a), social and ecological systems must not be studied in isolation, but as coupled social-ecological systems (Berkes and Folke, 1998). This is important to capture the potentially unexpected non-linear dynamics of self-reinforcing feedbacks between the ecological and the social sphere.

Conceptually, entering such a safe and just operating space means finding ways to overcome interlinked social and ecological dilemmas. A social dilemma is typically defined as a situation in which any individual prefers the socially defecting choice, regardless of what the other individuals choose. Yet, all individuals are better off if all choose the socially cooperative option (Dawes, 1980; Hardin, 1968). Ecological dilemmas are less uniformly defined. The choice of weighting current with expected future rewards (see e.g. Stern, 2008) presents an especially prominent ecological dilemma. Other ecological dilemmas have been shown to exist between the desirability, safety and flexibility of an outcome and management options (Heitzig et al., 2016) as well as the choice on an appropriate paradigm for environmental governance (Barfuss et al., 2018, P4; Heitzig et al., 2018, P3).

Social dilemmas have often been investigated using the framework of repeated normal form games (Axelrod and Hamilton, 1981; Bladon and Galla, 2011; Macy and Flache, 2002; Nowak, 2006; Szolnoki et al., 2009) proposing direct reciprocity in the form of repeated interactions as one way to foster the evolution of cooperation. However, the setting of repeated normal form games assumes a static environment and is therefore not suited to incorporate ecological dilemmas.

This chapter proposes the evolution of cooperation in stochastic games (Hilbe et al., 2018) as a framework for studying how coupled social-ecological dilemmas can be resolved. Stochastic games (Neyman and Sorin, 2003; Shapley, 1953) extend repeated normal form games by incorporating multiple environmental states. States can affect the agents' available actions, observations and current rewards. Transitions between states depend on chosen actions and generally occur probabilistically. In particular, Hilbe et al. (2018) use numerical simulations to find that agents require a sufficiently large farsightedness for cooperation to emerge in a stochastic game (c.f. Chapter 4, especially Fig. 4.9).

In this chapter, a particular stochastic game is introduced, termed the Ecological Public Good (EcoPG; see Fig. 5.1 and detailed description below). It extends the established public good by an environmental tipping element. Defection in the EcoPG is not only associated with the socially sub-optimal outcome but also with a probability to collapse the environment into a degraded state. In this degraded state agents can only receive a negative environmental impact payoff. From the degraded state only the cooperation action opens the chance to recover to the prosperous state.

Threshold public good games or likewise collective-risk social dilemmas are similar approaches to model catastrophic tipping (Barrett and Dannenberg, 2012; Barrett and Dannenberg, 2013; Dannenberg et al., 2014; Milinski et al., 2008; Tavoni et al., 2011; Vasconcelos et al., 2014; Walker and Gardner, 1992). Thus, through a behavioral experiment Milinski et al. (2008) find that cooperation requires a sufficiently large risk of collapse as well as a sufficiently large impact. However, in contrast to the EcoPG those games typically do not consider a recovery after the collapse has occurred. Yet, the EcoPG can been regarded as a generalization of these games. Letting the recovery probability parameter of the EcoPG approach zero resembles their situation. Further, the EcoPG is a comparable simple environment, using the framework of repeated interactions within a stochastic game. Extending environments within this framework allows the investigation of other, more complicated situations, as e.g. Leibo et al. (2017), Lindkvist and Norberg (2014), Lindkvist et al. (2017), and Pérolat et al. (2017) have done by using algorithmic reinforcement learning.

This chapter will investigate the EcoPG environment with the derived multi-state learning equations of Chapter 4. With respect to the emergence of cooperation a critical minimum farsightedness for each point in behavior phase space is derived, which is required for the agents to have to learn the cooperation agreement. Conversely, there may exist points in behavior space from which individual learners cannot reach the cooperation solution, since the farsightedness parameter is bounded. This situation describes one of three qualitatively different parameter regimes, that could be identified with respect to the parameters farsightedness, collapse risk and collapse damage.

The resolution of the Sustainable Development Goals and the Paris Agreement on climate change can be regarded as an agreement to cooperate towards a sustainable future. Yet in light of increasing political polarization (Dunlap et al., 2016) this chapter specifically asks for the stability of such a cooperation agreement, when one of the participating parties puts less weight on expected future gains. It can be shown that cooperation can remain stable despite considerable shortsightedness, but only if this agent's leverage to collapse the environment and the negative collapse impact are large. Conversely, this model projects that an actor, who does not believe in likely and severe consequences of a tipping catastrophe, will break off the cooperation agreement.

Sec. 5.2 introduces the Ecological Public Good and presents analytical results based on the learning equations derived in Chapter 4. Sec. 5.3 discusses the results before a summary closes this chapter in Sec. 5.4.

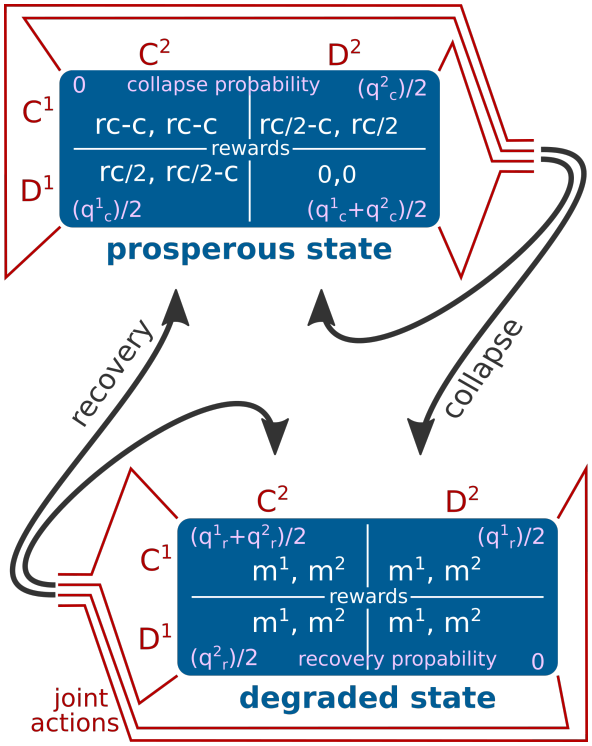

Figure 5.1: Ecological Public Good (shown for $N = 2$ agents) extends the repeated public good game to a stochastic game with two environmental states: In the prosperous state the agents play a standard public good game, in the degraded state agents have to endure an environmental collapse impact m^i. State transition depend on the joint actions, occur probabilistically and are visualized with black arrows. To simplify the presentation, it is assumed that collapse and recovery leverages $q_c^i \leq 1$ and $q_r^i \leq 1$, for both $i = 1, 2$.

5.2 Model and methods

5.2.1 The ecologcial public good

The Ecological Public Good (EcoPG) extends the public good normal form game to a stochastic game with two environmental states $s \in \mathcal{S} = \{\mathsf{p}, \mathsf{g}\}$: a prosperous and a degraded one. In each state s each agent $i \in \{1, \ldots, N\}$ can choose an action from its action set $a^i \in \mathcal{A}^i = \{\mathsf{c}, \mathsf{d}\}$: cooperation and defection. Fig. 5.1 illustrates the model for $N = 2$ agents.

Rewards $R^i_{sas'}$ follow the normal public good in the prosperous state:

$$R^i_{\mathsf{p}a^i\mathbf{a}^{-i}\mathsf{p}} = \begin{cases} rc\frac{N_\mathsf{c}}{N} - c & \text{if } a^i = \mathsf{c} \\ rc\frac{N_\mathsf{c}}{N} & \text{if } a^i = \mathsf{d} \end{cases} . \tag{5.1}$$

where c denotes the cost of cooperation, r the cooperation synergy factor and N_c the number of cooperating agents in the current time step. All cooperating agents contribute the cooperation cost to the public good, which gets multiplied by the synergy factor and is then distributed to all agents. Thus, for the special case of

$N = 2$ agents the payoff matrix reads

$$\begin{pmatrix} R^1_{\mathsf{pccp}} & R^1_{\mathsf{pcdp}} \\ R^1_{\mathsf{pdcp}} & R^1_{\mathsf{pddp}} \end{pmatrix} = \begin{pmatrix} rc - c & rc/2 - c \\ rc/2 & 0 \end{pmatrix} = \begin{pmatrix} r_r & r_s \\ r_t & r_p \end{pmatrix}. \tag{5.2}$$

Written in Prisoner's Dilemma form $r_r = rc - c$ denotes the cooperation reward, $r_t = rc/2$ the temptation reward, $r_s = rc/2 - c$ the sucker's reward and $r_p = 0$ the punishment reward. Likewise, the rewards for agent 2 in the prosperous state read

$$\begin{pmatrix} R^1_{\mathsf{pccp}} & R^2_{\mathsf{pcdp}} \\ R^1_{\mathsf{pdcp}} & R^2_{\mathsf{pddp}} \end{pmatrix} = \begin{pmatrix} rc - c & rc/2 \\ rc/2 - c & 0 \end{pmatrix} = \begin{pmatrix} r_r & r_t \\ r_s & r_p \end{pmatrix}. \tag{5.3}$$

However, when a state transition involves the degraded state $\mathbf{g}$, the agents only receive an environmental collapse impact m^i:

$$R^i_{\mathsf{pag}} = R^i_{\mathsf{gag}} = R^i_{\mathsf{gap}} = m^i, \quad \text{for all } \mathbf{a}. \tag{5.4}$$

State transitions from the prosperous to the degraded state occur with probability

$$T_{\mathsf{pag}} = \min\left(1, \sum_{i=1}^{N} \delta_{\mathsf{d}a^i} \frac{q^i_c}{N}\right), \tag{5.5}$$

where q^i_c is the leverage of agent i to collapse the environment under the defective action, expressed by the Kronecker delta $\delta_{\mathsf{d}a^i}$. The total collapse probability is maximal one or else the sum of individual collapse contributions. The leverage q^i_c is parameterized such that a value of $q^i_c = 1$ for all agents i would make a collapse exactly certain if all agents were defecting. For all joint actions $\mathbf{a}$ the probability to remain in the prosperous state must be $T_{\mathsf{pap}} = 1 - T_{\mathbf{pag}}$.

Similarly, for state transitions from the degraded to the prosperous state, a recovery leverage q^i_r is associated to the cooperative action. They give the total recovery probability according to

$$T_{\mathbf{gap}} = \min\left(1, \sum_{i=1}^{N} \delta_{\mathsf{c}a^i} \frac{q^i_r}{N}\right). \tag{5.6}$$

Hence, the probability to remain in the degraded state is given by $T_{\mathbf{gag}} = 1 - T_{\mathbf{gap}}$ for all joint actions $\mathbf{a}$.

5.2.2 Emergence of cooperation

First, the emergence of cooperation is investigated when agents individually learn through reinforcements according to the learning equations derived in Chapter 4. Specifically of interest are the preconditions for successful cooperation with respect to the agents' initial behavior, their farsightedness, collapse leverage and impact. To make the calculations feasible, the special case of $N = 2$ agents is considered.

As shown in Chapter 4, the relative temporal difference error arrows are pointing in the behavior phase space direction towards the next behavior the agents learn. Here, the ansatz is assumed that cooperation will be eventually obtained if these arrows are pointing more towards the cooperative solution than to the detective solution. Or in other words, if an initial change of behavior will increase overall cooperation probability, the agents will end up cooperating. Thus, this ansatz can be seen as an approximation of first order. How well it works remains to be investigated.

To obtain the hypersurface in behavior space dividing this cooperative from the defective basin of attraction following this ansatz, one has to search for behavior profiles $\mathbf{X}$ under which the temporal difference errors are pointing neither to the cooperative nor to the defective solution. Mathematically this can be expressed as

$$TD^1_{\mathsf{pc}}(\mathbf{X}) - TD^1_{\mathsf{pd}}(\mathbf{X}) \stackrel{!}{=} -\left(TD^2_{\mathsf{pc}}(\mathbf{X}) - TD^2_{\mathsf{pd}}(\mathbf{X})\right). \tag{5.7}$$

This equation is solved with the help of a computer algebra program (Meurer et al., 2017), since inserting the behavior profile $\mathbf{X}$, the reward and transition tensors, $\mathbf{R}$ and $\mathbf{T}$ into the temporal difference errors in Eq. 5.7 and subsequent terms within those temporal difference errors (see Chapter 4) yields an expression, unpractical to handle by hand. To make the calculation feasible only the prosperous state is considered, under the assumption that agents have a uniformly random behavior in the degraded state ($X^1_{\mathsf{gc}} = X^2_{\mathsf{gc}} = 0.5$). Additionally, the analysis is focused on Actor-Critic learning, or equivalently on SARSA learning under the assumption of a sufficiently large exploitation level β. Doing so eases the calculation because the forgetting term either does not exist (in the case of Actor-Critic learning, Eq. 4.14) or vanishes under $\beta \to \infty$ (in the case of SARSA learning, Eq. 4.12). Not considering the forgetting term here is justified because it favors neither the cooperative nor the defective solution. The forgetting term in the temporal difference error favors the center of behavior space ($X^1_{\mathsf{pc}} = X^2_{\mathsf{pc}} = 0.5$). The cooperative as well as the defective solution lie at edge ($X^1_{\mathsf{pc}} = X^2_{\mathsf{pc}} = 1$ or $X^1_{\mathsf{pc}} = X^2_{\mathsf{pc}} = 0$).

Using a computer algebra program one is able to solve Eq. 5.7 for X^2_{pc} as a function of X^1_{pc}, depending on the model's parameters (see Secs. 5.3). Assuming homogeneous agents with identical parameters Eq. 5.7 can be solved for e.g. the farsightedness γ (Sec. 5.3.1) or the environmental collapse impact m (Sec. 5.3.2). Agent-indices of the parameters will be omitted for homogeneous agents. Likewise, wherever model parameters occur with omitted agent-indices parameter-homogeneous agents are considered.

5.2.3 Stability of cooperation

Having reached the cooperative solution, now its stability is investigated, assuming heterogeneous agents. Specifically, the preconditions for stable cooperation with respect to a single agent's farsightedness, collapse leverage and impact are of interest. Without loss of generality, the heterogeneous agent is denoted by agent 1. To obtain the analytical solution for the edge of the cooperation region in the parameter space of agent 1 the following ansatz is assumed: cooperation must be as attractive as defection to the agent in the propsperous state:

$$TD^1_{\mathsf{pc}}({}^{\mathsf{c}}\mathbf{X}) \stackrel{!}{=} TD^1_{\mathsf{pd}}({}^{\mathsf{c}}\mathbf{X}) \tag{5.8}$$

with the behavior profile ${}^{\mathsf{c}}X^i_{s\mathsf{c}} \to 1$, for all i, s, since the agents are close to the cooperative solution in all states.

As above, the focus is put on Actor-Critic learning or SARSA learning with a sufficiently large level of exploitation β^1. The remaining terms in the temporal difference errors (Eqs. 4.12 and 4.14, respectively) left to calculate are the reward ${}_{\mathbf{TX}^{-i}}\langle R\rangle^1_{sa}\big|_{\mathbf{X}={}^{\mathsf{c}}\mathbf{X}}$ and the value of the next state ${}^{\text{next}}\mathcal{V}^1_{sa}({}^{\mathsf{c}}\mathbf{X})$. To simplify the presentation, it is assumed that collapse and recovery leverages are not exceeding the value 1: $(q^1_r + q^2_r)/2 \leq 1$ and $q^1_c \leq 1$.

Inserting reward and transition tensors as given above in Prisoner's Dilemma form into the behavioral and temporal average ${}_{\mathbf{TX}^{-i}}\langle ... \rangle$ as defined in Eqs. 3.5 and 3.6 yields

$$\left.\begin{pmatrix} {}_{\mathbf{TX}^{-i}}\langle R\rangle^1_{\mathsf{pc}} & {}_{\mathbf{TX}^{-i}}\langle R\rangle^1_{\mathsf{pd}} \\ {}_{\mathbf{TX}^{-i}}\langle R\rangle^1_{\mathsf{gc}} & {}_{\mathbf{TX}^{-i}}\langle R\rangle^1_{\mathsf{gd}} \end{pmatrix}\right|_{\mathbf{X}={}^{\mathsf{c}}\mathbf{X}} = \begin{pmatrix} r_r & \frac{q^1_c}{2}m^1 + \left(1 - \frac{q^1_c}{2}\right) r_t \\ m^1 & m^1 \end{pmatrix}. \tag{5.9}$$

Likewise, using the reward and transition tensor as given above in Prisoner's Dilemma form to compute the next value according to Eq. 4.13 yields

$$\begin{pmatrix} {}^{\text{next}}\mathcal{V}^1_{\mathsf{pc}}({}^{\mathsf{c}}\mathbf{X}) & {}^{\text{next}}\mathcal{V}^1_{\mathsf{pd}}({}^{\mathsf{c}}\mathbf{X}) \\ {}^{\text{next}}\mathcal{V}^1_{\mathsf{gc}}({}^{\mathsf{c}}\mathbf{X}) & {}^{\text{next}}\mathcal{V}^1_{\mathsf{gd}}({}^{\mathsf{c}}\mathbf{X}) \end{pmatrix} = \begin{pmatrix} V^1_{\mathsf{p}}({}^{\mathsf{c}}\mathbf{X}) & \frac{q^1_c}{2}V^1_{\mathsf{g}}({}^{\mathsf{c}}\mathbf{X}) + \left(1 - \frac{q^1_c}{2}\right) V^1_{\mathsf{p}}({}^{\mathsf{c}}\mathbf{X}) \\ \frac{q^1_r+q^2_r}{2}V^1_{\mathsf{p}}({}^{\mathsf{c}}\mathbf{X}) + \left(1 - \frac{q^1_r+q^2_r}{2}\right) V^1_{\mathsf{g}}({}^{\mathsf{c}}\mathbf{X}) & \frac{q^2_r}{2}V^1_{\mathsf{p}}({}^{\mathsf{c}}\mathbf{X}) + \left(1 - \frac{q^2_r}{2}\right) V^1_{\mathsf{g}}({}^{\mathsf{c}}\mathbf{X}) \end{pmatrix}, \tag{5.10}$$

with the state values

$$\begin{pmatrix} V_{\mathsf{p}}^1({}^{\mathsf{c}}\mathbf{X}) \\ \\ V_{\mathsf{g}}^1({}^{\mathsf{c}}\mathbf{X}) \end{pmatrix} = \begin{pmatrix} r_r \\ \frac{\frac{q_r^1+q_r^2}{2}\gamma^1 r_r + (1-\gamma^1)m^1}{1-\left(1-\frac{q_r^1+q_r^2}{2}\right)\gamma^1} \end{pmatrix}. \tag{5.11}$$

Inserting these results into the ansatz expressed in Eq. 5.8 one obtains implicitly after rearranging

$$\boxed{\hat{r}_t - \hat{r}_r = \frac{\hat{q}_c^1}{2}\left[\hat{r}_t - \hat{m}^1 + \frac{\hat{\gamma}^1}{1-\hat{\gamma}^1(1-\hat{p}_r^{\mathrm{tot}})}(\hat{r}_r - \hat{m}^1)\right]} \tag{5.12}$$

with $\hat{p}_r^{\mathrm{tot}} = (\hat{q}_r^1 + \hat{q}_r^2)/2$. Interestingly, the only parameter of agent 2 occurring in Eq. 5.12 is its recovery leverage q_r^2. This is because the learners are independent, in the sense that they do not take into account expected future actions of other agents. Thus, Eq. 5.12 describes agent 1's deliberations when it is equally prone to remain in and to break off the cooperation agreement. It describes the edge of the stable cooperation regime in the parameter space of agent 1. The reward difference between the temptation and cooperation reward $(r_t - r_r)$ must equal the collapse probability in the case of agent 1's defection $(q_c^1/2)$ multiplied by the reward difference between the temptation reward and the collapse impact $r_t - m^1$ plus the discounted expected future reward difference between the cooperation reward and the collapse impact $\frac{\gamma^1}{1-\gamma^1(1-p_r^{\mathrm{tot}})}(r_r - m^1)$. Eq. 5.12 nicely demonstrates that there is no distinct reward value and that only reward differences are of interest to the agent.

Note that the stability concept presented here results directly from the dynamical perspective brought forth by the self-learning agents. A comparison to related concepts, such as classic game theoretic equilibria in stochastic games (see e.g. Neyman and Sorin, 2003) may be of interested for future work, is however beyond the scope and focus of this chapter.

5.3 Discussion of results

5.3.1 Critical farsightedness

First, the dependence of the emergence of cooperation on the farsightedness γ for parameter-homogeneous agents is considered. In the degraded state rewards are incidental for all actions. Only the cooperation action will eventually lead back to the prosperous state with higher rewards. Thus, agents easily learn to cooperate in the degraded state (Fig. 5.2 a-c).

In the prosperous state, however, it depends on the initial behavior profile and the farsightedness γ of the agents, whether or not they will learn to cooperate or defect. From the ansatz expressed in Eq. 5.7 one can compute the critical discount factor γ_{crit}

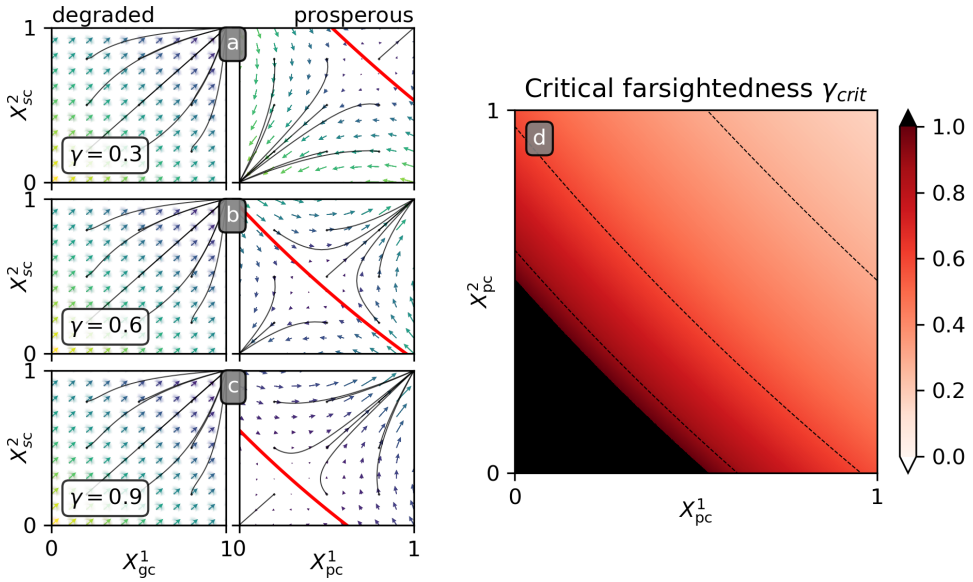

Figure 5.2: Critical farsightedness in behavior space. (a-c) The arrows indicate the average direction the learning dynamics drive the parameter-homogeneous agents towards for each state, averaged over all behavior space points of the other state for different farsightedness γ. The algebraic solutions where these arrows have x,y components of equal length but different sign are shown in red. This ansatz is able to divide the behavior space in a cooperation and a defection basin of attraction, as exemplary trajectories, shown in gray, indicate. Remaining parameters are $r = 1.2, c = 5, m = -1.5$, $q_c = 0.8$, $q_r = 0.1$. The critical farsightedness γ_{crit} from this ansatz is shown in the behavior space section of the prosperous state (d). From an arbitrary point in behavior space both agents need a farsightedness of at least γ_{crit} to learn the cooperative solution. Thus, from the black area there is no such farsightedness since γ is confined by 1.

in behavior space above which agents will learn to cooperate (Fig. 5.2 d). Fig. 5.2 a-c nicely confirms this ansatz. The red lines resulting from Eq. 5.7 are indeed capable to separate the behavior space into a cooperative and a defective basin of attraction as individual learning trajectories in gray show. This result is in line with previous findings that cooperation needs a sufficiently large farsightedness (Hilbe et al., 2018). Yet, with the approach taken here it is possible to assign a critical farsightedness γ_{crit} to every point in behavior space. Only with farsightedness parameters equal or greater than this critical value, cooperation will emerge. Interestingly, Fig. 5.2 d shows an area (in black) in behavior space from which the cooperation solution cannot be obtained by the individual reinforcement learners, regardless of their farsightedness $\gamma \in [0, 1)$.

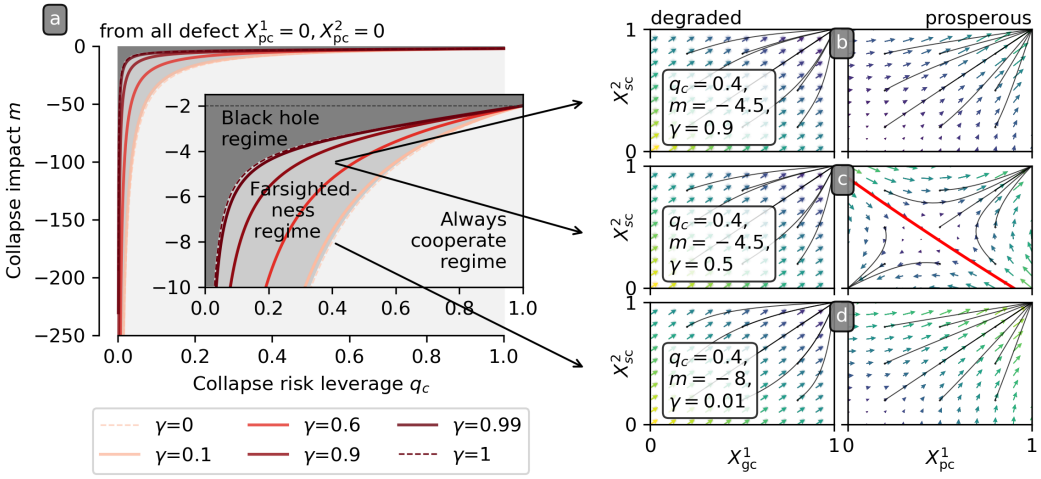

Figure 5.3: Three regimes for the learning of cooperation from full defection. (a) For parameter-homogeneous agents the ansatz Eq. 5.7 , solved for the environmental collapse impact parameter m is plotted against the collapse risk leverage q_c for different values of farsightedness γ. The behavior profile was set to full defection in the prosperous state ($X^1_{pc} = X^2_{pc} = 0$). Thus, in environments with damage parameters (m, q_c) below the curves, agents with a farsightedness γ at least as large as denoted by the curves, will learn to cooperate from an all defect solution. Behavior space sections are shown for three parameter combinations of q_c, m, γ (b-d). The arrows indicate the average direction the learning dynamics drive the agents towards for each state, averaged over all behavior space points of the other state. Similar as in Fig. 5.2 a-c, the ansatz Eq. 5.7 is shown in behavior space as red lines. Example learning trajectories are shown in gray. Remaining parameters are $r = 1.2, c = 5, q_r = 0.1$.

5.3.2 Learning to cooperate from full defection

Under what conditions does such a, metaphorically speaking, black hole for the learning dynamics exit? It does not exist, if agents learn to cooperate even from an initial behavior profile close to the all defect point in the behavior space of the prosperous state. The ansatz described in Sec 5.2.2 with the all defect initial conditions $X^1_{pc} = X^2_{pc} = 0$ is solved for the environmental collapse impact parameter m and plotted against the collapse risk leverage q_c for different discount factors γ (Fig. 5.3 a). Thus, damage parameter combinations (m, q_c) right from or below the curves shown in Fig. 5.3 a designate environmental conditions in which agents with a farsightedness γ of at least as large as marked by the respective curve will learn to fully cooperate, even from an initial all defect behavior in the prosperous state. Conversely, in environments with damage parameter combinations (m, q_c) left from or above these curves, the learning agents with less farsightedness as marked by the respective curve, will remain at the all defect behavior. Fig. 5.3 b-d confirms this relation by showing exemplary behavior phase spaces for three parameter combinations. Fig. 5.3 b and c only differ in their discount factor. The environmental damage point $(q_c = 0.4, m = -4.5)$ lies between the curves associated to discount factors $\gamma = 0.6$ and $\gamma = 0.9$. In such environments the ansatz expressed

in Eq. 5.7 predicts that agents with a farsightedness of at least $\gamma = 0.9$ learn to cooperate even from the all defect solution, as shown in Fig. 5.3 a. Fig. 5.3 b nicely confirms this prediction. Yet, agents with a farsightedness of $\gamma < 0.6$ are predicted to not learn to cooperate from full defection, as Fig. 5.3 c confirms as well. They might, however, learn to cooperate from more cooperative initial conditions.

Interestingly, the bounded parameter $\gamma \in [0, 1)$ divides this environmental damage parameter space into three regimes: i) Left from or above the dashed curve in dark red, marking $\gamma = 1$, is the area in which such a black hole, as shown in Fig. 5.2, exists. Even agents with close to absolute farsightedness will not escape the all defect behavior. ii) The area enclosed by the two dashed lines describes environmental damage parameter combinations (m, q_c) for which there exist discount factors γ, such that the agents will learn to cooperate from the initial all defect behavior. iii) Right from or below the dashed curve in light right, marking $\gamma = 0$, is the area for which agents learn to cooperate from the defective behavior, regardless of the discount factor. This is true even for discount factors $\gamma \to 0$, as Fig. 5.3 d nicely confirms. Agents will learn to cooperate regardless of their farsightedness, if the risk through the leverage to collapse the environment q_c and the collapse impact in the degraded state m are both sufficiently large.

Fig. 5.3 a suggests that the relationship between collapse impact and risk leverage is dominated by an $m \propto 1/q_c$ term. Bringing q_c on the other side yields the expected collapse impact $q_c m$, influencing the agents' learning. This result is in line with previous findings from experimental threshold public good games, reporting that a large expected damaging impact is beneficial for cooperation (Milinski et al., 2008). Yet, through the approach taken here it is possible to derive analytical dependencies between the involved parameters, based on the assumption of individual learning agents.

As $q_c = 1$, the curves in Fig. 5.3 a do not depend on γ. This is intuitively explainable. If immediate collapse is certain under the defective action, the agents do not require farsightedness. Conversely, for decreasing collapse leverages q_c, the γ-dependence increases.

With respect to the range of collapse impact values, Fig. 5.3 a shows that above a value of $m = -2$, i.e. less negative impact in the case of collapse, there exists always a black hole regime, regardless of the collapse leverage q_c. This value $m = -2$ is exactly equal to the sucker's payoff $r_s = rc/2 - c = -2$ for the parameters as chosen in Fig. 5.3. Thus, if the equally distributed collapse impact m is slightly better than the unequally distributed suffering due to exploitation r_s, i.e. $m > r_s$, agents close to the defective behavior will not learn to cooperate, even if collapse is certain.

5.3.3 Stability of cooperation

After examining the preconditions for the emergence of cooperation, this section now discusses its stability. Specifically of interest here is the stability of cooperation when one agent changes its characteristics, e.g. its attitude towards expected future rewards or its expected impact in the case of collapse. In other words, this section

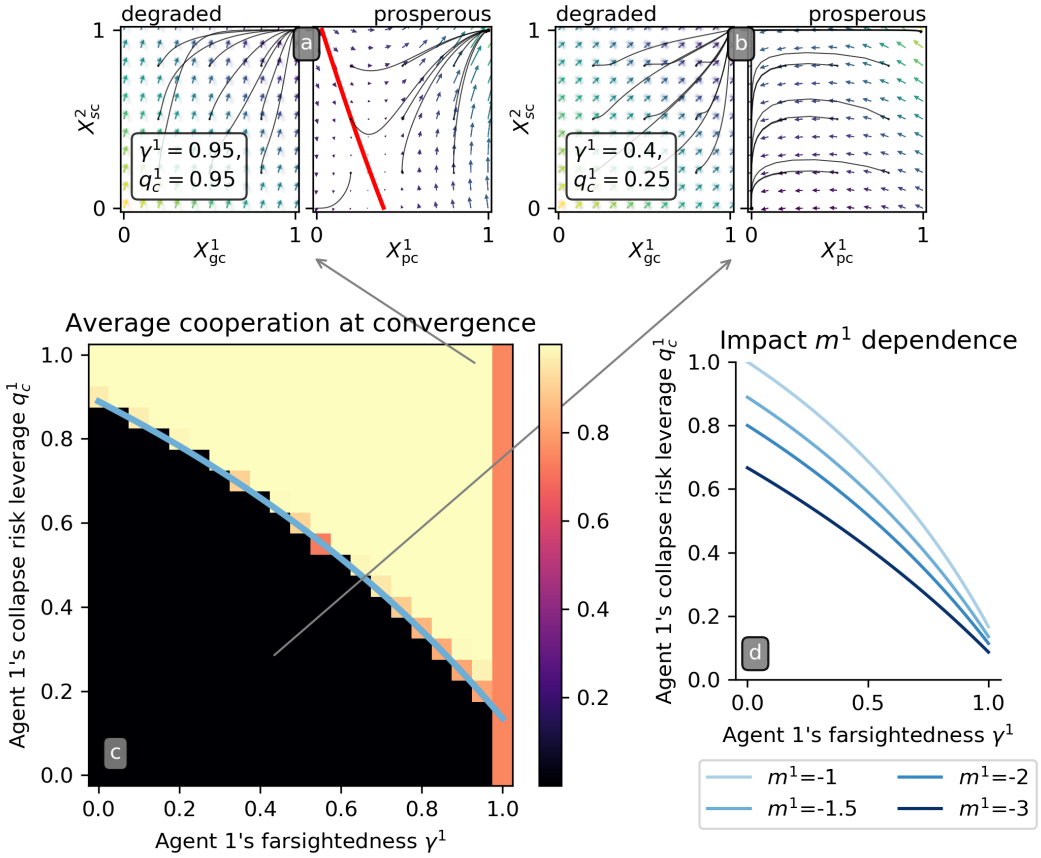

Figure 5.4: Stability of cooperation for heterogeneous agents. The parameter space shown (c) is spanned by the farsightedness γ^1 and the collapse propensity q_c^1 of agent 1. The color indicates the level of cooperation after the learning has converged from an initial level of cooperation of $X_{sc}^i = 0.99$ for all agents ($i = 1, 2$) and all states ($s = \mathrm{p}, \mathrm{g}$). The analytical solution is shown in blue (Eq. 5.12), which is in good accordance with the numerical learning. The behavior spaces of two parameter examples are shown (a,b), one from the cooperation stable region, one from the cooperation brake off region. The analytical solution for different levels of environmental collapse impact m^1 of agent 1 shows that a larger impact $-m^1$ is beneficial for cooperation (d). Remaining parameters are $r = 1.2, c = 5, m^1 = m^2 = -1.5$, $q_c^2 = 0.8$, $q_r^1 = q_r^2 = 0.1$, $\gamma^2 = 0.8$, if not specified otherwise.

highlights the parameter space of agent 1 with respect to the parameter combinations under which the cooperation solution remains stable, or under which combinations agent 1 breaks off the cooperation agreement. As shown with Eq. 5.12, the stability of cooperation depends only on agent 1's discount factor γ^1, collapse risk leverage q_c^1 and collapse impact m^1 and the overall recovery probability, and does not depend on the respective parameters of the other agent.

Depending on the these parameters of agent 1, cooperation either remains stable or breaks down (Fig. 5.4). Fig. 5.4a,b shows how the dynamics in behavior space become asymmetric with respect to the angle bisector due to the agents' heterogeneity. The red lines result from the same ansatz as in Fig. 5.2, separating the cooperative from the defective basin of attraction. Fig. 5.4 c compares numerical simulations with the analytical solution of the stability separating line in parameter space (Eq. 5.12). They are in good accordance, justifying the assumptions made in Sec. 5.2.3. Fig. 5.4 c shows that cooperation can remain stable despite agent 1 having comparable small farsightedness. Even if agent 1 is not caring for the future at all ($\gamma^1 = 0$), cooperation can remain stable. In this case the stability diving line in parameter space (Eq. 5.12) reduces to $\hat{r}_t - \hat{r}_r = \hat{q}_c^1/2 \cdot [\hat{r}_t - \hat{m}^1]$. Cooperation remains stable if the expected reward difference between the temptation reward r_t and the environmental impact m^1 in the case of collapse ($q_c^1/2$) is greater than the reward difference between the temptation r_t and the cooperation reward r_r. Thus, for large collapse leverages q_c^1 and likewise large damage impacts $-m^1$ cooperation remains stable, even if the agent's farsightedness $\gamma^1 = 0$. Fig. 5.4 d confirms for all γ^i that the greater the suffering the agent has to endure in the case of collapse the larger the stable region in its parameter space. On the other hand Eq. 5.12 suggests that if the collapse leverage $q_c^1 = 0$ there is no parameter combination that keeps the cooperation agreement stable. This is because Eq. 5.12 reduces to $\hat{r}_t - \hat{r}_r = 0$ in this case. Yet, from the definition of a social dilemma $r_t > r_r$ must hold. Thus, $r_t - r_r \leq 0$ cannot be fulfilled leaving the cooperation solution unstable. This result is intuitively explainable, because if there is no collapse risk leverage $q_c^1 = 0$, at least for agent 1, the game reduces to a classic social dilemma with a dominating defection solution.

To summarize, an individual learning agent with low collapse leverage or likewise low expected damaging impact is more likely to break off the cooperation agreement because in its deliberations it comes to the conclusion that its own actions will not have sufficiently severe consequences, and therefore defection seems the favorable choice.

5.4 Summary

This chapter introduced a novel environment, termed the Ecological Public Good. It extends the established social dilemma public good game by an environmental tipping element. This environment thereby represents a stylized coupled social-ecological dilemma.

The emergence and stability of cooperation in the EcoPG were analyzed by applying the individual reinforcement learning equations derived in Chapter 4. Specifically, this chapter focused on the relationship between two previously unconnected characteristics: the agents' farsightedness, weighting current with expected future rewards for a learning behavior update; and the environment's expected damage to the agents in the case of collapse. For each characteristic individually, previous theoretical or empirical findings could be reproduced.

Overall, this chapter demonstrates the usefulness of agent-environment interface model designs for advancing the mathematics of sustainability (Levin, 2013). Using the derived learning equations in Chapter 4, it was possible to present analytical results for the preconditions of both, the emergence and stability of cooperation.

With respect to the emergence of cooperation, computer-algebraic analytical results were shown, assigning a critical farsightedness to each point in behavior space, above which cooperation emerges. Further, the existence of a *black hole* regime in behavior space was demonstrated, from which there is no farsightedness, such that the agents would be able to learn the cooperative solution. Agents prefer to collectively suffer in environmental collapse rather than cooperating in a prosperous environment. This is an interlinked social-ecological dilemma, presenting an interesting challenge for future research: Is it possible to design non-trivial learning equations, that find the cooperative solution from this regime?

It was shown that such a black hole exits regardless of the collapse risk if the impact due to environmental collapse is less severe than the suffering due to being exploited in the social dilemma.

Concerning the parameter relationship between the risk to collapse the environment, the level of environmental impact in the case of collapse, and the farsightness of the agents, three regimes with respect to the learning of cooperation could be identified: i) an always cooperate regime, in which risk and damage are so large, that the agents will learn to cooperate, regardless of their farsightedness; ii) a farsightedness regime, where it depends on the farsightedness of the agents, whether they learn to cooperate from a full defection initial behavior; and iii) the black hole regime, where these agents are not able to learn to cooperate from full defection, even if they have full farsightedness.

With respect to the stability of cooperation for heterogeneous agents, i.e. when one agent changes its parameters, it was shown through analytical and numerical calculations that an individual learner can keep the cooperation solution stable despite considerable shortsightedness. However, this is only the case if its leverage to collapse the environment and the expected damage in the case of collapse are large. Since cooperation can remain stable even when the agent has no farsightedness at all, but not if the collapse leverage vanishes, one may conclude, that the expected collapse damage due to the actors own action is of greater importance than the actor's farsightedness. Thus, an actor who does not believe in likely and severe consequences of a tipping catastrophe will break off the cooperation agreement.

Outlook. The learning agents used in this chapter are individual incremental reward-maximizing learners. Thus, they do not take into account any expected future actions of other agents. The effect of such more complicated learners (c.f Busoniu et al., 2008) on both, the emergence and stability of cooperation, is of great interest for future work. For example, it has been shown in a related setting that higher levels of strategic reasoning can make cooperation agreements unstable (Verendel et al., 2015).

Likewise of great interest is the relevance of the agents' reward-maximizing paradigm within the learning equations. Optimization approaches for environmental governance have been criticized for delivering short-term gains at the expense of long-term environmental degradation. This is why the next chapter will perform a systematic comparison of a reward optimization decision paradigm with two other important decision paradigms for environmental governance: sustainability and the safe operating space.

Python code for the reproduction of the reported results is available upon request by the author.

Chapter 6

Third act: Decision paradigms for the governance of tipping elements

Jeder muß bereit sein, sich einsperren und wirtschaftlich ruinieren zu lassen; wenn sich genug Personen finden, die diesen harten Weg gehen, werden wir Erfolg haben.

Bulle 1 - from Paul Dessau's *Einstein*: Third act

The learning equations derived in Chapter 4 follow a reward-*optimization* paradigm. Yet, optimization approaches for environmental governance have been criticized for delivering short-term gains at the expense of long-term environmental degradation. It is thus of importance to asses the effects of such decision paradigms for the preconditions to enter sustainable pathways. Prominent alternative paradigms to derive behaviors or likewise policies from are *sustainability* and the *safe operating space*.

This chapter will systematically compare these three decision paradigms for the management of a tipping element environment, which can be regarded as the special case of a single agent in the Ecological Public Good, as introduced in Chapter 5. Since the reinforcement learning dynamics result directly from the reward-optimization approach, there are no equivalent learning dynamics for the other paradigms. Therefore, no learning is used in this chapter and policies (equivalent to behaviors) are derived directly from the respective decision paradigms.

It can be shown that optimization can lead to sustainable and safe policies but is by no means guaranteed to do so. In fact, no paradigm guarantees fulfilling requirements imposed by another paradigm. This chapter presents simple heuristics for the conditions under which these trade-offs occur. Further, it demonstrates that the absence of such a master paradigm is of special relevance for governing the climate system, which may reside at edge in the parameter space where economic optimization becomes neither sustainable nor safe.

This chapter is based on (Barfuss et al., 2018, P4).

6.1 Introduction

The Sustainable Development Goals (UN General Assembly, 2015) and the adoption of the Paris Agreement on climate change (COP, 2015) set the target of prosperous development for people and the planet. Yet, it remains challenging to translate these aims into concrete policy implementations, accounting for non-linearities, such as tipping elements (Lenton et al., 2008; Schellnhuber, 2009), regime shifts (Lade et al., 2013; Scheffer et al., 2001), and multi-stabilities (Donges et al., 2017) as well as multiple kinds of uncertainties (Anderies et al., 2007; Irwin et al., 2016; Polasky et al., 2011a), and extreme events (Farmer et al., 2015).

Optimization approaches have emerged as the primary guiding principle to derive a policy strategy for environmental governance (Perman et al., 2003; Weyant, 2014). Most often, the present value of macroeconomic social welfare, i.e. the sum of discounted future benefits minus costs, is the target to be optimized. Such optimization approaches have been criticized regarding the discount rates used, delivering short term gains at the expense of long-term environmental degradation (Ackerman et al., 2009; Stern, 2008). Further criticism targets the lack of a systems perspective required to understand the structural landscape of model dynamics, as well as the assumptions made due to imperfect information (Donges et al., 2017; Farmer et al., 2015; Irwin et al., 2016). This critique is partly dealt with in optimization variants, such as robust (Anderies et al., 2007; Woodward and Tomberlin, 2014) or viable (De Lara and Doyen, 2008; Martinet and Doyen, 2007; Rougé et al., 2013) control, which are dealing with multiple types of uncertainty (Chadès et al., 2016). Naturally, other or multiple objectives (Branke et al., 2008) and criteria (Ehrgott, 2000; Greco et al., 2016) with possible constraints (Altman, 1999) can be optimized as well. In this work, the term is used solely in the narrow economic sense of maximizing the present value as defined in Eq. 6.9 below.

Sustainability. In recognition of increasing environmental and social threats (Meadows et al., 1972) the policy paradigm of sustainability has emerged in the scientific and political discourse (Pezzey, 1992; WCED, 1987). The economics of sustainability has brought up many definitions of sustainability alone (Arrow et al., 2012; Fleurbaey, 2015; Gerlagh, 2017; Pezzey, 1997). In these analyses sustainability is usually imposed as a constraint within an economic welfare optimization paradigm. Trade-offs to economic welfare optimization are well known (Pezzey, 1997, 2004). However, these classic social welfare optimization approaches are challenged through the increasing recognition of non-linearities, such as tipping points, regime shifts, uncertainties and the risk of catastrophic outcomes (Donges et al., 2017; Irwin et al., 2016). Taking up these challenges, e.g. non-convexities (Dasgupta and Mäler, 2004) and climate tipping elements (Cai et al., 2016; Lontzek et al., 2015) have been studied within an economic framework. In this chapter, the formal definition of sustainability is derived from the Brundtland report (WCED, 1987). Its design is deliberately

simple and targeted to the mathematical framework used (see below). This definition is not intended to be applicable to a general model of a welfare economy (Perman et al., 2003; Pezzey, 1992).

Safe operating space. Recent advances in sustainability science have brought forth tolerable windows (Petschel-Held et al., 1999) or safe operating spaces (Dearing et al., 2014; Rockström et al., 2009a) as a policy paradigm to derive concrete actions from (Carpenter et al., 2015). These concepts originate from resilience thinking (Folke et al., 2010) and a precautionary principle (Raffensperger and Tickner, 1999) to deal with potential dangerous tipping elements in the environmental governance system. Trade-offs but also synergies with optimization thinking have been discussed (Fischer et al., 2009). Also formal analyses studying relations between resilience as a system property and sustainability were conducted (Derissen et al., 2011; Mäler and Li, 2010).

Research challenge. However, the reciprocal relationships between these three paradigms of economic optimization, sustainability and safe operating space is still insufficiently explored. Such an understanding is important in order to judge, for example, when economic optimization is, or is not, an appropriate policy goal. Also, guidance is required when a sustainability paradigm may conflict with a safe operating space paradigm and vice versa.

Overview. In this chapter progress is reported towards a better understanding of the mutual relationships between these three paradigms of economic optimization, sustainability and safe operating space by applying them to a single-agent tipping element environment. This is done because of the increasing importance of tipping points and regime shifts in environmental governance. The agent-environment interface model is deliberately stylized, thereby applicable across multiple cases and scales, to gain a deeper understanding more complex models might miss. The formal definitions of the three paradigms are designed to fit the mathematical framework (see below). Since there is no focus on intragenerational justice in this chapter, one agent suffices as a decision making subject, in contrast to the multi-agent setting of e.g. Chapter 5.

Sec. 6.2 describes the agent-environment tipping element, presents the formal definitions of the three investigated governance paradigms and derives analytical expressions for the paradigms classification of two relevant governance policies in the parameter space of the model. Sec. 6.3 presents and discusses these results including a qualitative comparison with the three real-world systems of climate, fisheries and farming before this chapter concludes with a summary in Sec. 6.4

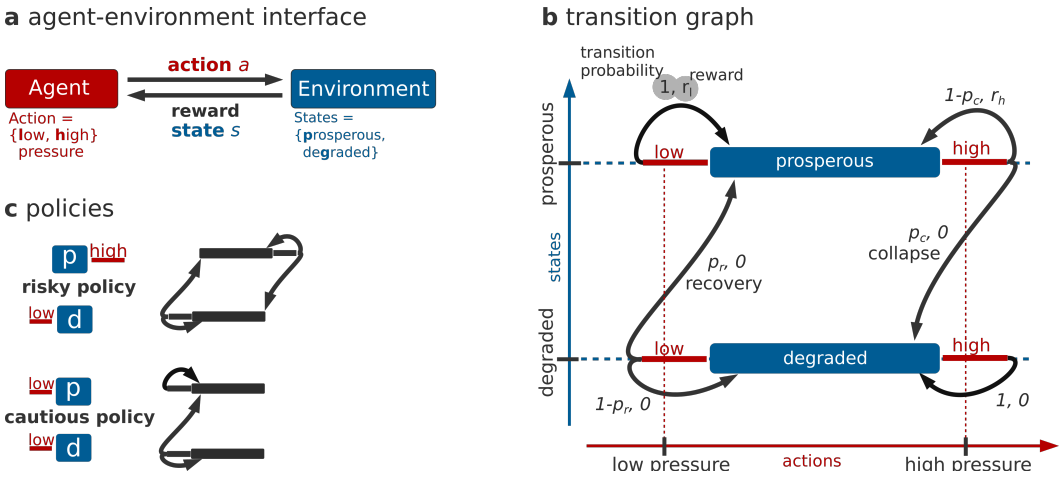

Figure 6.1: Single-agent tipping element environment. (a) Agent-environment interface: based on the state information and received reward, the agent chooses an action a from its actions set to gain rewards. (b) The transition graph gives state transition probabilities and corresponding rewards for all triples of state s, action a, next state s', i.e. in state s the agent takes action a and moves to state s'. (c) Risky and cautious policies including the resulting Markov chains as a transition graph.

6.2 Model and methods

6.2.1 A single-agent tipping element environment

The special case of one agent ($N = 1$) reduces the multi-agent environment system, as presented in Chapter 3 to a Markov decision processes (Bellman, 1957; Puterman, 2005), in which one agent makes decisions about how to interact with its environment (Fig. 6.1 a). Like in the Ecological Public Good (EcoPG, Chapter 5), the particular environment of this chapter can reside in either a prosperous state p, which provides immediate rewards to the agent, or a degraded state g, from which the agent receives no reward, i.e.

$$\mathcal{S} = \{\mathsf{p}, \mathsf{g}\}. \tag{6.1}$$

At each time step, the agent chooses between two actions a, exerting either a high or low pressure on the environment, i.e.

$$\mathcal{A} = \{\mathsf{h}, \mathsf{l}\}. \tag{6.2}$$

Since there is only one agent, the EcoPG's action names cooperation and defection are potentially confusing. Also, agent identifiers in the superscript can be omitted.

At the prosperous state, taking the low pressure action the agent is guaranteed to receive reward r_l and remain at the prosperous state:

$$R_{\mathsf{plp}} = r_\mathsf{l}; \quad T_{\mathsf{plp}} = 1. \tag{6.3}$$

However, the agent faces the dilemma that taking the high pressure action, it may receive reward r_h (which is typically larger than r_l), but risks triggering a collapse of the environment to the degraded system state with non-zero probability p_c and no immediate reward at all:

$$R_{\mathsf{php}} = r_\mathrm{h}, \quad T_{\mathsf{php}} = 1 - p_c, \tag{6.4}$$
$$R_{\mathsf{phg}} = 0, \quad T_{\mathsf{phg}} = p_c. \tag{6.5}$$

From there, only the low pressure action opens the option to recover to the prosperous state with non-zero probability p_r,

$$R_{\mathsf{glp}} = 0, \quad T_{\mathsf{glp}} = p_r, \tag{6.6}$$
$$R_{\mathsf{glg}} = 0, \quad T_{\mathsf{glg}} = 1 - p_r, \tag{6.7}$$
$$R_{\mathsf{ghg}} = 0, \quad T_{\mathsf{ghg}} = 1. \tag{6.8}$$

For example, the high pressure action could correspond to a technological optimistic policy, emitting a business-as-usual amount of carbon to the atmosphere which yields a reward of high, short-term economic output as long as the system has not tipped. The low pressure action resembles emitting a reduced amount of carbon, assuming a lower short-term economic output for the guarantee to not trigger climate tipping elements into a disastrous state.

The agent chooses its action according to its policy $\mathbf{X}$. Since this model is about environmental governance $\mathbf{X}$ is referred to by policy instead of behavior. However, mathematically $\mathbf{X}$ is used identically as introduced in Chapter 3.

Similarly as defined in Eq. 3.9 the value $V_s(\mathbf{X})$ of a state s under a given policy $\mathbf{X}$ is given by the expected value of the normalized accumulated discounted rewards $r(t)$ with discount factor or farsightedness $0 \leq \gamma \leq 1$ when starting in state $s(0) = s$ and following policy $\mathbf{X}$:

$$V_s(\mathbf{X}) = {}_{\mathbf{TX}}\left\langle \lim_{T\to\infty} \frac{\sum_{t=0}^{T} \gamma^t r(t)}{\sum_{t=0}^{T} \gamma^t} \;\middle|\; s(0) = s \right\rangle_s . \tag{6.9}$$

6.2.2 Derivation of value functions

In the following section analytical expressions of the value functions (Eq. 6.9) are derived when applied to the Markov decision process as described above. These are needed for the definitions of the three governance paradigms.

There are four deterministic policies in this Markov decision process model:

1. the *risky* policy (${}^rX_{\mathsf{ph}} = 1, {}^rX_{\mathsf{gl}} = 1$),
   applying the high pressure action at the prosperous state and the low pressure action at the degraded state;

2. the *cautious* policy (${}^cX_{\mathsf{pl}} = 1, {}^cX_{\mathsf{gl}} = 1$),
   applying the low pressure action at the prosperous and the degraded state;

3. the third policy (${}^3X_{\mathsf{ph}} = 1, {}^3X_{\mathsf{dh}} = 1$),
   applying the high pressure action at the prosperous and the degraded state; and
4. the fourth policy (${}^4X_{\mathsf{pl}} = 1, {}^4X_{\mathsf{gh}} = 1$),
   applying the low pressure action at the prosperous state and the high pressure action at the degraded state.

It is sufficient to concentrate on deterministic policies, because if an optimal policy exits there exits also a deterministic optimal policy (Puterman, 2005). Further, the focus is put on the first two policies only, which were named the risky and the cautious policy (Fig. 6.1 c) . The remaining two policies apply a high pressure action at the degraded state, which will trap the agent at this position for eternity without receiving any reward. The mathematics for these policies is left to the interested reader.

In the following, the analytical expressions of the state values of these policies will be derived as functions of the parameters $(p_c, p_r, \gamma, r_\mathrm{l}, r_\mathrm{h})$. Readers who are not interested in the mathematical details may safely skip this section.

Case $\gamma < 1$

From Eq. 6.9 and for $\gamma < 1$ one can derive the recursive relationship between state values, known as the Bellman equation (Bellman, 1957; see also Eq. 3.14):

$$V_s(\mathbf{X}) = \sum_a \sum_{s'} X_{sa} T_{sas'} \left[(1-\gamma) R_{sas'} + \gamma V_{s'}(\mathbf{X})\right]. \tag{6.10}$$

Applied to the tipping model the value for the prosperous state reads

$$V_{\mathsf{p}}(\mathbf{X}) = \begin{cases} p_c \gamma V_{\mathsf{g}}(\mathbf{X}) + (1-p_c)[(1-\gamma) r_\mathrm{h} + \gamma V_{\mathsf{p}}(\mathbf{X})] & \text{for } X_{\mathsf{ph}} = 1 \\ (1-\gamma) r_\mathrm{l} + \gamma V_{\mathsf{p}}(\mathbf{X}) & \text{for } X_{\mathsf{pl}} = 1 \end{cases}. \tag{6.11}$$

The value for the degraded state is given by

$$V_{\mathsf{g}}(\mathbf{X}) = \begin{cases} \gamma V_{\mathsf{g}}(\mathbf{X}) & \text{for } X_{\mathsf{gh}} = 1 \\ (1-p_r)\gamma V_{\mathsf{g}}(\mathbf{X}) + p_r \gamma V_{\mathsf{p}}(\mathbf{X}) & \text{for } X_{\mathsf{gl}} = 1 \end{cases}. \tag{6.12}$$

To obtain the explicit state values for the risky policy (${}^r X_{\mathsf{ph}} = 1, {}^r X_{\mathsf{gl}} = 1$) one needs to solve the system of equations

$$V_{\mathsf{p}}({}^r\mathbf{X}) = p_c \gamma V_{\mathsf{g}}({}^r\mathbf{X}) + (1-p_c)\left[(1-\gamma) r_\mathrm{h} + \gamma V_{\mathsf{p}}({}^r\mathbf{X})\right] \tag{6.13}$$

$$V_{\mathsf{g}}({}^r\mathbf{X}) = (1-p_r)\gamma V_{\mathsf{g}}({}^r\mathbf{X}) + p_r \gamma V_{\mathsf{p}}({}^r\mathbf{X}), \tag{6.14}$$

which yields

$$V_{\mathsf{p}}({}^{r}\mathbf{X}) = r_{\mathrm{h}} \frac{(1-p_c)(1-(1-p_r)\gamma)}{1-(1-p_c-p_r)\gamma} \tag{6.15}$$

$$V_{\mathsf{g}}({}^{r}\mathbf{X}) = r_{\mathrm{h}} \frac{(1-p_c)p_r\gamma}{1-(1-p_c-p_r)\gamma}. \tag{6.16}$$

To obtain the explicit state values for the cautious policy (${}^{c}X_{\mathsf{pl}} = 1, {}^{c}X_{\mathsf{gl}} = 1$) one needs to solve the system of equations

$$V_{\mathsf{p}}({}^{c}\mathbf{X}) = (1-\gamma)r_{\mathrm{l}} + \gamma V_{\mathsf{p}}({}^{c}\mathbf{X}) \tag{6.17}$$

$$V_{\mathsf{g}}({}^{c}\mathbf{X}) = (1-p_r)\gamma V_{\mathsf{g}}({}^{c}\mathbf{X}) + p_r\gamma V_{\mathsf{p}}({}^{c}\mathbf{X}), \tag{6.18}$$

which yields

$$V_{\mathsf{p}}({}^{c}\mathbf{X}) = r_{\mathrm{l}} \tag{6.19}$$

$$V_{\mathsf{g}}({}^{c}\mathbf{X}) = r_{\mathrm{l}} \frac{p_r\gamma}{1-(1-p_r)\gamma}. \tag{6.20}$$

Case $\gamma = 1$

For $\gamma = 1$ the values $V(\mathbf{X})$ become independent from the initial state. One can compute them by multiplying the stationary state $\boldsymbol{\sigma}(X)$ of the effective Markov chain transition matrix ${}_{\mathbf{X}}\langle T\rangle_{ss'}$ with the reward vector ${}_{\mathbf{TX}}\langle \mathbf{R}\rangle$:

$$V(\mathbf{X}) = \boldsymbol{\sigma}(\mathbf{X}) \cdot {}_{\mathbf{TX}}\langle \mathbf{R}\rangle \,. \tag{6.21}$$

Encoding the state vector with the prosperous state in the first and the degraded state in the second dimension, the effective Markov transition matrix for the risky policy reads

$${}_{\mathbf{X}}\langle \mathbf{T}\rangle|_{\mathbf{X}={}^{r}\mathbf{X}} = \begin{pmatrix} {}_{\mathbf{X}}\langle T\rangle_{\mathsf{pp}} & {}_{\mathbf{X}}\langle T\rangle_{\mathsf{pg}} \\ {}_{\mathbf{X}}\langle T\rangle_{\mathsf{gp}} & {}_{\mathbf{X}}\langle T\rangle_{\mathsf{gg}} \end{pmatrix}\Bigg|_{\mathbf{X}={}^{r}\mathbf{X}} = \begin{pmatrix} 1-p_c & p_c \\ p_r & 1-p_r \end{pmatrix}. \tag{6.22}$$

Its left eigenvector with eigenevalue one reads

$$\boldsymbol{\sigma}({}^{r}\mathbf{X}) = \begin{pmatrix} \sigma_{\mathsf{p}}({}^{r}\mathbf{X}) & \sigma_{\mathsf{g}}({}^{r}\mathbf{X}) \end{pmatrix} = \frac{1}{p_c+p_r} \begin{pmatrix} p_r & p_c \end{pmatrix}. \tag{6.23}$$

Thus, for the risky policy the ratio of residence times (of the prosperous state over the degraded state) equals the ratio of transition probabilities (of recovery over collapse).

The reward receivable from each state reads

$$ {}_{\mathbf{TX}}\langle \mathbf{R} \rangle|_{\mathbf{X}={}^{r}\mathbf{X}} = \left(\begin{array}{c} {}_{\mathbf{TX}}\langle R \rangle_{\mathsf{p}} \\ \\ {}_{\mathbf{TX}}\langle R \rangle_{\mathsf{g}} \end{array} \right) \Bigg|_{\mathbf{X}={}^{r}\mathbf{X}} = \left(\begin{array}{c} (1-p_c) r_{\mathrm{h}} \\ \\ 0 \end{array} \right) . \tag{6.24} $$

As one can easily see, evaluating $V(\mathbf{X}) = \boldsymbol{\sigma}(\mathbf{X}) \cdot {}_{\mathbf{TX}}\langle \mathbf{R} \rangle$ for the risky policy yields consistent results with the calculation for $0 \leq \gamma < 1$ from above: The value $V(\mathbf{X})$ for the case $\gamma = 1$ is identical to inserting $\gamma = 1$ into Eqs. 6.15 or 6.16.

For the cautious policy the effective Markov transition matrix reads

$$ {}_{\mathbf{X}}\langle \mathbf{T} \rangle|_{\mathbf{X}={}^{c}\mathbf{X}} = \left(\begin{array}{cc} {}_{\mathbf{X}}\langle T \rangle_{\mathsf{pp}} & {}_{\mathbf{X}}\langle T \rangle_{\mathsf{pg}} \\ \\ {}_{\mathbf{X}}\langle T \rangle_{\mathsf{gp}} & {}_{\mathbf{X}}\langle T \rangle_{\mathsf{gg}} \end{array} \right) \Bigg|_{\mathbf{X}={}^{c}\mathbf{X}} = \left(\begin{array}{cc} 1 & 0 \\ \\ p_r & 1-p_r \end{array} \right) . \tag{6.25} $$

Its left eigenvector with eigenevalue one reads

$$ \boldsymbol{\sigma}({}^{c}\mathbf{X}) = \left(\sigma_{\mathsf{p}}({}^{c}\mathbf{X}) \;\; \sigma_{\mathsf{g}}({}^{c}\mathbf{X}) \right) = \left(1 \;\; 0 \right) . \tag{6.26} $$

As intuitively obvious, the agent remains in the prosperous state under the cautious policy. The reward receivable from each state reads

$$ {}_{\mathbf{TX}}\langle \mathbf{R} \rangle|_{\mathbf{X}={}^{c}\mathbf{X}} = \left(\begin{array}{c} {}_{\mathbf{TX}}\langle R \rangle_{\mathsf{p}} \\ \\ {}_{\mathbf{TX}}\langle R \rangle_{\mathsf{g}} \end{array} \right) \Bigg|_{\mathbf{X}={}^{c}\mathbf{X}} = \left(\begin{array}{c} r_{\mathrm{l}} \\ \\ 0 \end{array} \right) . \tag{6.27} $$

As with the risky policy, the value $V(\mathbf{X})$ for the cautious policy for the case $\gamma = 1$ is consistent with the calculations for $0 \leq \gamma < 1$ from above (Eqs. 6.19 or 6.20).

6.2.3 Paradigm definitions

These policies may be classified according to whether they are economic welfare optimal or not, sustainable or not, and safe or not, as defined as follows:

Optimal

The definition for optimality is taken from a standard textbook:

> A policy $\mathbf{X}$ is defined as *optimal* (in the economic welfare sense) if its value $V_s(\mathbf{X})$ (Eq. 6.9) for every state s is larger than or equal to the value of any other policy (Puterman, 2005).

Sustainable

Based on the Brundtland commission's report on sustainable development (WCED, 1987) a sustainable policy should fulfill two requirements: First, meet the needs of the present. This gets formally translated into the agent evaluating the present state s as acceptable (similar to viable (Martinet and Doyen, 2007), tolerable (Petschel-Held et al., 1999) or desirable (Heitzig et al., 2016)), if its value (Eq. 6.9) exceeds a normatively chosen minimum acceptable value $r_{\min}$:

$$s \text{ acceptable under } \mathbf{X} \text{ iff } \; V_s(\mathbf{X}) \geq r_{\min} \tag{6.28}$$

Note that the division of state space into acceptable and unacceptable states is not identical for all polices, but depends on the rewards receivable through executing a policy. Second, a sustainable policy should sustain the ability to meet the needs of the future (WCED, 1987).

Thus, a policy $\mathbf{X}$ is hereby defined as *sustainable* if every state the agent eventually visits under policy $\mathbf{X}$ is acceptable (Eq. 6.28).

Note that this reduction of sustainability to the one-dimensional value $V_s(\mathbf{X})$ has much similarity with the notion of weak sustainability (Neumayer, 2003).

Safe

The Safe Operating Space (SOS; Rockström et al., 2009a) is typically defined as a subset of the whole state space $\mathcal{S}$, containing favorable system states bounded by thresholds (Carpenter et al., 2015; Steffen et al., 2015a). In practice, the position of these potential tipping thresholds is always uncertain and the boundaries are placed at the lower end of the uncertainty zone. In that way the definition of the states within the safe operating space constitutes a normative judgment about the risk the decision maker is willing to tolerate. In the subsequent analyses the extreme position of no risk tolerance will be taken and the SOS is identified with only the (more favorable) prosperous state, independent of the collapse probability p_c.

A policy $\mathbf{X}$ is hereby defined as *safe* if every state the agents eventually visits under policy $\mathbf{X}$ lies within the SOS.

In contrast to acceptable and unacceptable states, SOS states are independent of the policy used.

6.2.4 Analytical expressions for paradigms classification of policies

This section presents analytical results which tell whether the risky and the cautious policy are optimal or not, sustainable or not and safe or not depending on the model

parameters $(p_c, p_r, \gamma, r_\mathrm{l}/r_\mathrm{h}, r_\mathrm{min}/r_\mathrm{h})$. Since all three rewards come in arbitrary units, the policy classification only depends on their ratios. Readers not interested in the mathematical details may safely skip this section.

Optimal

To derive the analytical expression of the hypersurface in parameter space that separates the regions where either the risky or the cautious policy is optimal, one has to set $V_\mathsf{p}({}^r\mathbf{X}) \stackrel{!}{=} V_\mathsf{p}({}^c\mathbf{X})$ (or equivalently $V_\mathsf{g}({}^r\mathbf{X}) \stackrel{!}{=} V_\mathsf{g}({}^c\mathbf{X})$, since the parameter combination where a policy is optimal is independent from the state) and implicitly obtains

$$\hat{r}_\mathrm{h} \cdot (1-\hat{p}_c)(1-\hat{\gamma}(1-\hat{p}_r)) = \hat{r}_\mathrm{l} \cdot (1-\hat{\gamma}(1-\hat{p}_c-\hat{p}_r)). \tag{6.29}$$

Sustainable

To obtain the hypersurface that separates state s being acceptable from being not acceptable under policy $\mathbf{X}$ one must apply the definition from Eq. 6.28: $V_s(\mathbf{X}) \stackrel{!}{=} r_\mathrm{min}$. Hence, for the risky policy at the prosperous state one can set $V_\mathsf{p}({}^r\mathbf{X}) \stackrel{!}{=} r_\mathrm{min}$ and obtains implicitly

$$\hat{r}_\mathrm{h} \cdot (1-\hat{p}_c))(1-\hat{\gamma}(1-\hat{p}_r)) = \hat{r}_\mathrm{min} \cdot (1-\hat{\gamma}(1-\hat{p}_c-\hat{p}_r)). \tag{6.30}$$

For the risky policy at the degraded state one can set $V_\mathsf{g}({}^r\mathbf{X}) \stackrel{!}{=} r_\mathrm{min}$ and obtains implicitly

$$\hat{r}_\mathrm{h} \cdot (1-\hat{p}_c)\hat{p}_r\hat{\gamma} = \hat{r}_\mathrm{min} \cdot (1-\hat{\gamma}(1-\hat{p}_c-\hat{p}_r)). \tag{6.31}$$

For the cautious policy at the prosperous state one can set $V_\mathsf{p}({}^c\mathbf{X}) \stackrel{!}{=} r_\mathrm{min}$ and obtains implicitly

$$\hat{r}_\mathrm{l} = \hat{r}_\mathrm{min}. \tag{6.32}$$

For the cautious policy at the degraded state one can set $V_\mathsf{g}({}^c\mathbf{X}) \stackrel{!}{=} r_\mathrm{min}$ and obtains implicitly

$$\hat{r}_\mathrm{l} \cdot \hat{p}_r\hat{\gamma} = \hat{r}_\mathrm{min} \cdot (1-\hat{\gamma}(1-\hat{p}_r)). \tag{6.33}$$

To get from acceptability to sustainability for the risky policy one has to logically combine Eqs. 6.30 and 6.31. The risky policy is sustainable only if both the prosperous and the degraded state are acceptable since it will visit both states recurrently. The safe policy is sustainable exactly where the prosperous state is acceptable since it will eventually end up and remain at the prosperous state. Fig. 6.2 shows an example of the acceptability division of state-parameter space and the resulting sustainability division.

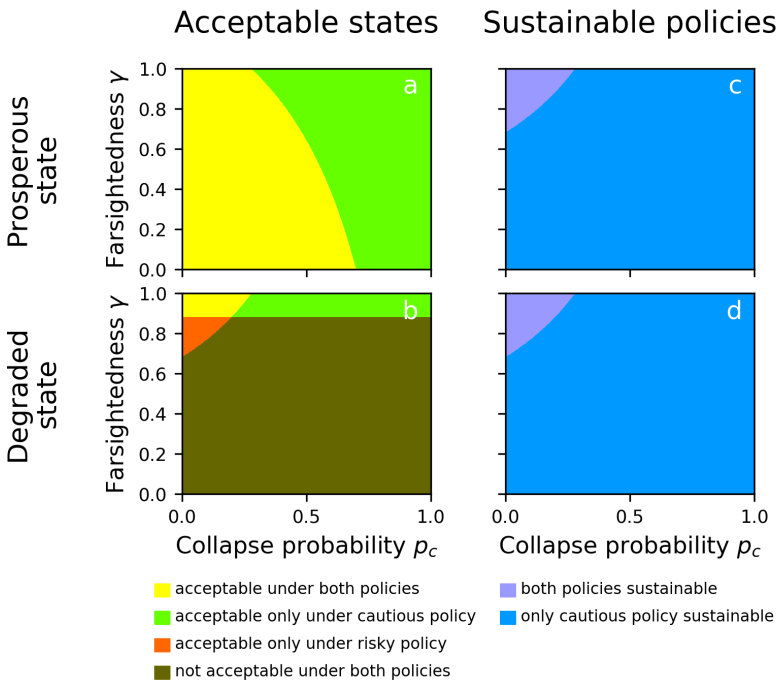

Figure 6.2: Sustainable policies are based on acceptable states as illustrated here in the parameter space (shown as collapse probability p_c vs. farsightedness γ with the prosperous state (a,c) and the degraded state (b,d)). Parameters are $p_r = 0.2$, $r_\mathrm{l}/r_\mathrm{h} = 0.5, r_\mathrm{min}/r_\mathrm{h} = 0.3$. In (a,b) color indicates whether the respective state is acceptable under different policies. In (c,d) color indicates the resulting sustainable policies.

Safe

The division of the parameter space according to the safe operating space paradigm is obvious from its definition. Only the cautious policy is a safe policy since it will eventually end up and remain in the prosperous, safe operating space state. The risky policy switches recurrently between the prosperous and the degraded state which makes it, by definition, not safe.

6.3 Discussion of results

In summary, the presented model of an agent-environmental tipping element depends on the five parameters $p_c, p_r, \gamma, r_\mathrm{l}/r_\mathrm{h}, r_\mathrm{min}/r_\mathrm{h}$: the probability of a collapse from the prosperous to the degraded state under the high pressure action p_c, the probability of recovery from the degraded to the prosperous state under the low pressure action p_r, the agent's farsightedness γ, the high reward receivable from the high pressure action when staying at the prosperous state r_h, the low reward receivable by taking the low pressure action at the prosperous state r_l, and the normatively chosen minimum

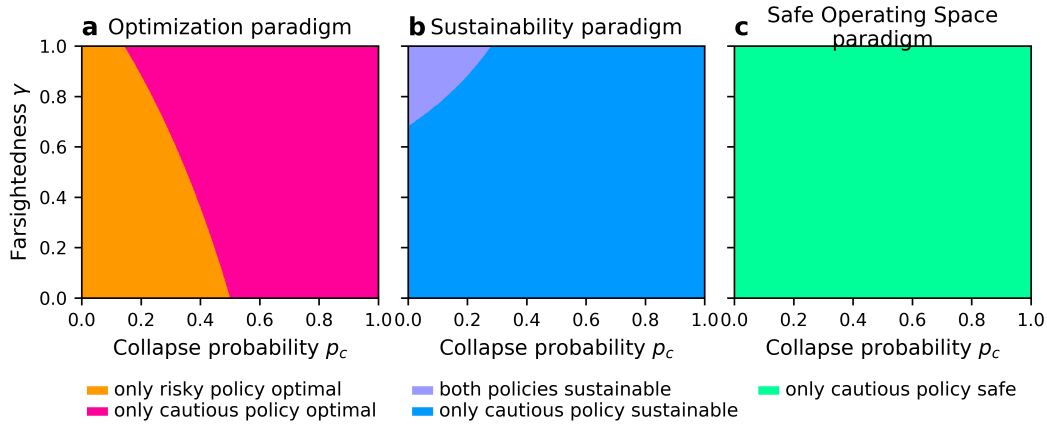

Figure 6.3: Classification of the risky and cautious policy according to the three policy paradigms: (a) optimization, (b) sustainability and (c) safe operating space in the model parameter space (shown here as collapse probability p_c vs. farsightedness γ); remaining parameters were chosen as $p_r = 0.2, r_\mathrm{l}/r_\mathrm{h} = 0.5, r_\mathrm{min}/r_\mathrm{h} = 0.3$ for illustration purposes. Colored regions result from analytically derived equations (see Sec. 6.2.4). Depending on the parameter region, both risky and cautious policy can be optimal and sustainable. Only the cautious policy is safe.

acceptable reward r_min a state value must have to be perceived as acceptable under a certain policy. Since all three rewards come in arbitrary units, the policy classification only depends on their ratios.

6.3.1 Classification of risky and cautious policy

Fig. 6.3 shows parameter regions in which the risky and the cautious policy are optimal or not, sustainable or not and safe or not in the parameter space section spanned by the discount factor γ and the collapse probability p_c, using the derived hypersurfaces (Eqs. 6.29 – 6.33). One can observe that above a certain critical value of the collapse probability p_c the cautious policy becomes optimal (Fig. 6.3 a, pink), despite the smaller immediate reward $r_\mathrm{l} = 0.5 r_\mathrm{h}$. This result confirms previous findings on optimal management with regime shifts (Polasky et al., 2011b).

Further, one finds a decreasing critical collapse probability with increasing farsightedness γ. Hence, for more farsighted agents the risky policy is optimal only for small collapse probabilities p_c (orange).

Provided the low pressure reward exceeds the normative minimum acceptable value threshold, $r_\mathrm{l} \geq r_\mathrm{min}$, then the cautious policy is sustainable for all parameter combinations $p_c, p_r, \gamma, r_\mathrm{l}/r_\mathrm{h}$ (Fig. 6.3 b, blue and purple). Only for small collapse probabilities p_c and simultaneously high farsightedness γ the risky policy becomes sustainable as well (purple). This is because in this parameter region the risky policy is acceptable also at the degraded state.

The cautious policy is a safe policy independently from the parameter combinations $p_c, p_r, \gamma, r_\mathrm{l}/r_\mathrm{h}, r_\mathrm{min}/r_\mathrm{h}$ (Fig. 6.3 c, green). There is no combination of parameters at which the risky policy is safe.

6.3.2 Relationships between paradigms

Which paradigm should a policy maker choose?, one might ask. Is there a paradigm superior to the other ones? Yet, one finds that policies can carry all logical combinations of the three examined paradigms (optimization, sustainability, safe operating space) in the parameter space of the presented model. In other words, policies can be all combinations of optimal or not, sustainable or not and safe or not. This yields a classification of policies into eight different categories (Fig. 6.4). Among them are four to be discussed in more detail:

opt & not sus. In particular, optimal policies are not necessarily sustainable (Fig. 6.4, red and yellow). This is the case if the normative value threshold r_min is too large. The cautious policy does not return enough value to be sustainable ($r_\mathrm{l} < r_\mathrm{min}$, yellow) and the risky policy at the degraded state produces too little future reward to be sustainable, due to the low chance of recovery and lack of farsightedness.

opt & not safe. Nor are optimal policies necessarily safe (Fig. 6.4, red and purple). This occurs in parameter regions where the risky policy is optimal. The risky policy cannot be safe because of the risk of collapse to the degraded state.

safe & not sus. A safe policy does not necessarily imply a sustainable policy either (Fig. 6.4, green and yellow). When the normative threshold value for sustainability r_min exceeds the reward from a low pressure action r_l: $r_\mathrm{min} > r_\mathrm{l}$, then the cautious policy is safe but not sustainable. Following a similar line of argument, the SOS concept(Rockström et al., 2009a) has been extended to a Safe And Just Operating Space (SAJOS) which additionally accounts for social indicators (Raworth, 2017), such as the number of people living in extreme poverty. Thus, SAJOS policies can be interpreted as the overlap of safe with sustainable policies. Within the presented model, one can give a definite criterion for when this form of SAJOS exists: as long as the reward from a low pressure action r_l exceeds the normative threshold value r_min ($r_\mathrm{l} > r_\mathrm{min}$), the cautious policy is both safe and sustainable (Fig. 6.4, cyan and gray).

sus & not safe. However, there exist also sustainable policies outside the SOS (Fig. 6.4, blue and purple.) These are risky policies (hence, not safe) with simultaneously high farsightedness γ and low collapse probability p_c. At those parameter regions the degraded state is still evaluated as acceptable due to sufficient anticipated future rewards and therefore the risky policy is sustainable.

The circumstance that parameter regimes exist that are sustainable but not safe and vice versa clearly stems from our definition of sustainability which resembles

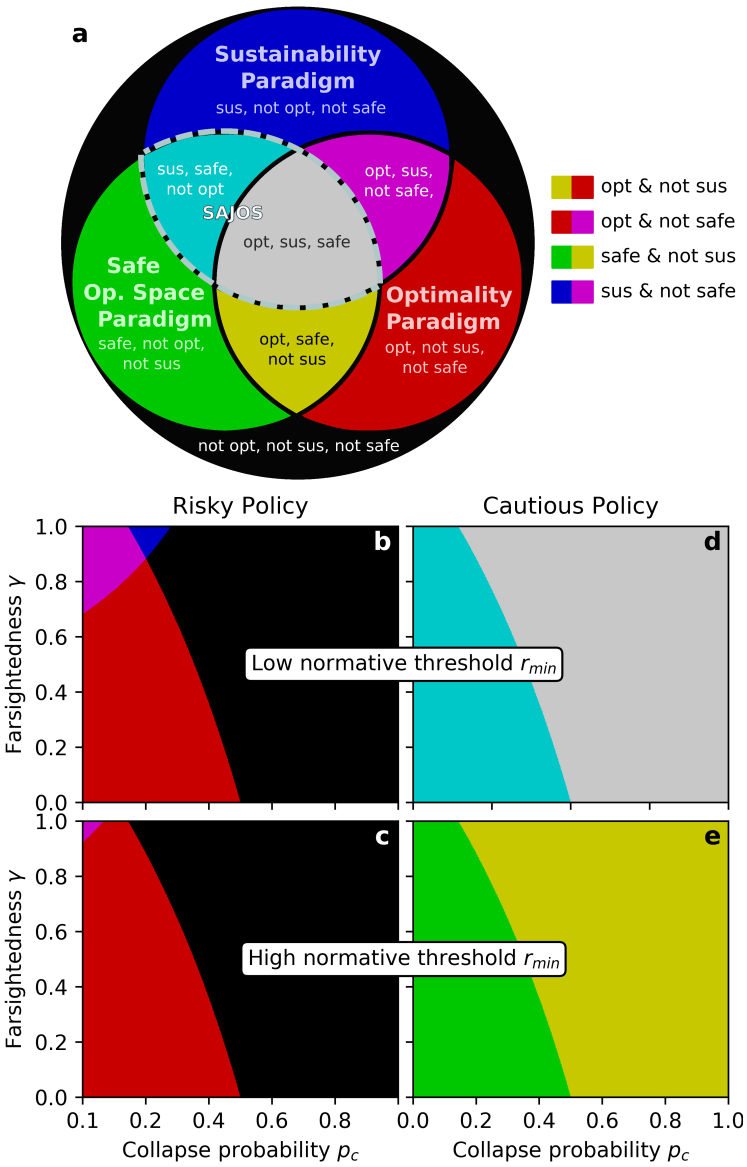

Figure 6.4: Paradigms combinations for risky and cautious policy. There exist policies in parameter space of this model for all logical combinations of paradigm classifications (a), i.e. a policy can be any combination of (not) optimal, (not) sustainable and (not) safe. Remaining parameters where chosen as $p_r = 0.2, r_\mathrm{l}/r_\mathrm{h} = 0.5$ for illustration purposes. For a sufficiently low normative threshold value $r_\mathrm{min} \leq r_\mathrm{l}$ (here $r_\mathrm{min}/r_\mathrm{h} = 0.3$) a Safe And Just Operating Space (SAJOS) exists, which was identified as the overlap of safe and sustainable policies (b,d) (gray and cyan area). For a sufficiently large $r_\mathrm{min} > r_\mathrm{l}$ (here $r_\mathrm{min}/r_\mathrm{h} = 0.7$) a SAJOS does not exist (c,e).

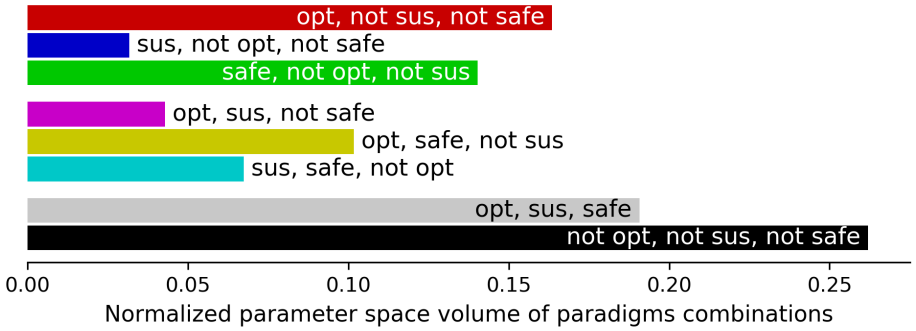

Figure 6.5: Fraction of parameter space volumes for all eight paradigms combination. All parameters ($p_c, p_r, \gamma, r_\mathrm{l}/r_\mathrm{h}, r_\mathrm{min}/r_\mathrm{h}$) were chosen linearly between 0 and 1 for both the risky and the cautious policy. As a direct consequence of the definitions of the safe operating space paradigm and the cautious and risky policy, all paradigm combinations which are safe correspond to the use of the cautious policy, in all others the risky policy was applied. A random decision making agent within a random tipping element will most likely end up with a policy that is neither optimal, neither sustainable nor safe, followed by the parameter sweet spot regime where the policy is simultaneously optimal, sustainable and safe. Interestingly, the third likeliest option is a parameter regime which is optimal, but neither sustainable nor safe.

a form of weak sustainability (Neumayer, 2003). By doing so one can conceptually separate the issues of environmental safety from social justice without compromising the target of a safe and just parameter space regime.

Note that this classification into the eight different policy paradigm combinations also applies to the case of absolute farsightedness ($\gamma = 1$; see the tops of Fig. 6.4 b-e). Thus, the trade-offs between the examined paradigms do not vanish, as one might presume considering the debate about appropriate discount rates (Nordhaus, 2007; Stern, 2008).

6.3.3 Volume of paradigm combinations

The previous section demonstrated that there exists no master paradigm among the three examined within the presented model. Hence, policies may carry all eight logical combinations of (not) optimal, (not) sustainable and (not) safe.

This section asks how large these eight regimes of paradigm combinations are in the whole parameter space. Fig. 6.5 shows the largest regime to be the combination that is neither optimal, neither sustainable nor safe followed by the parameter sweet spot regime in which all paradigms yield the cautious policy as optimal, sustainable and safe. Together they constitute a parameter space volume of approx. 45% in which the three paradigms of economic optimization, sustainability and safe operating space align with each other in yielding the same policy. Interestingly, the third likeliest option is the paradigm combination in which the risky policy is optimal but neither sustainable nor safe. This is the most likeliest parameter regime among those where

the paradigms yield different policies. Thus, blindly applying economic optimization in this stylized tipping element has a significant chance of leading to policies that are neither sustainable nor safe.

On the other hand, the volume of the safe and just operating space (gray and cyan bars in Fig. 6.5) is comparable to the most likeliest (black) regime. Thus, about one out of four random decision making agents interacting with a random tipping element will end up in the safe and just operating space.

6.3.4 Application to real-world human-environment tipping elements

The above policy classification offers valuable insights for the governance of real-world human-environment systems. This section discusses how this analysis relates to the cases of the climate system, fisheries and farming. The purpose of this discussion is to gain a qualitative understanding how the presented model relates to important real-world challenges of environmental governance, not a detailed assessment of the latter.

Therefore, the respective collapse and recovery probabilities per time step p_c and p_r of the model are estimated via the typical timescales on which these systems remain in one state or the other (see below). A model time step is mapped to a year. Let p be the probability per time step that a system state will transition into another state. The average number of time steps the system will be in that state is given by $\langle D \rangle = (1-p)/p$. Inverting yields $p = 1/(\langle D \rangle + 1)$. Thus, a collapse timescale of e.g. 50 years corresponds to a collapse probability of $p_c \approx 0.02$. Tab. 6.1 summarizes the assumed transition timescales and corresponding transition probabilities.

	Climate	Fishery	Farming
Assumed collapse timescale [years]	~ 40	~ 20	~ 100
Corresponding collapse probability	~ 0.025	~ 0.045	~ 0.01
Assumed recovery timescale [years]	$\to\sim \infty$	~ 50	~ 300
Corresponding recovery probability	$\to\sim 0$	~ 0.02	~ 0.003

Table 6.1: Typical transition timescales and corresponding probabilities.

Additionally, a parameter sensitivity analysis is added by visualizing the likelihood of ending up in a certain parameter regime by color gradients between regimes (Fig. 6.6).

Climate system. Regarding the climate system, one has to acknowledge that several interacting tipping elements contribute to the system's behavior (Lenton et al., 2008) and its representation as a single tipping element is a huge simplification on its own. Nevertheless, it is assumed that the current state of the climate system is still comparable to the prosperous one of our model. If planetary thresholds are

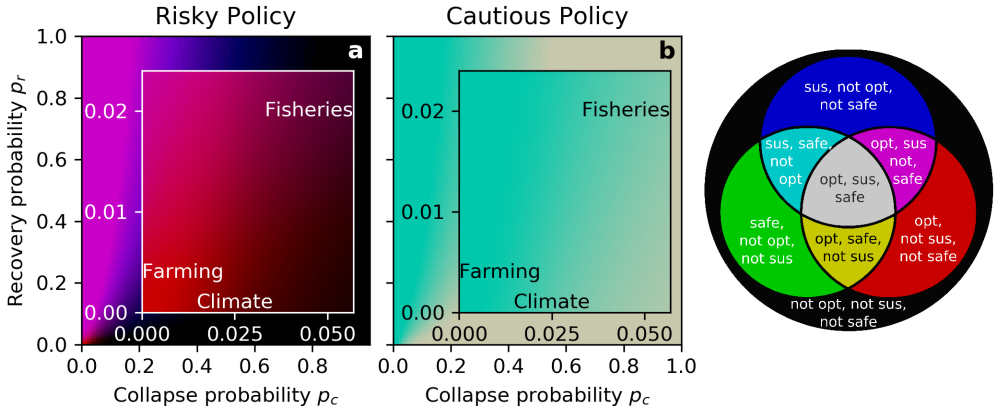

Figure 6.6: Human-environment systems in paradigms combinations for risky and cautious policy here shown in model parameter space of collapse probability p_c versus recovery probability p_r. Color indicates the paradigms combination similarly as in Fig. 6.4. Here, additional gradual changes between the color regimes indicate the probability of being in a certain paradigms combinations regime under parameter uncertainty ranges. Remaining parameters where chosen linearly within the range of $0.95 \leq \gamma \leq 0.99$, $0.3 \leq r_\mathrm{l}/r_\mathrm{h} \leq 0.7, 0.1 \leq r_\mathrm{min}/r_\mathrm{h} \leq 0.5$. The approx. transition probabilities p_c and p_r were assigned to the human-environment systems climate, fisheries and farming according to the typical timescale these systems spent in one state. For farming, a risky policy is likely to be optimal but neither sustainable nor safe. The parameter uncertainty of the other parameters does not allow a clear statement in which parameter regime fisheries are likely to fall. The climate system may lie at the edge of the sweet spot, where all paradigms yield the cautious policy. However, for a smaller collapse probability p_c optimization is more likely to yield the risky policy, which becomes also neither sustainable nor safe at this point.

crossed, there is the risk to enter a Hothouse Earth state (Steffen et al., 2018), in analogy to the model's degraded state. Relevant timescales for triggering a collapse lie within 30 to 50 years under a business-as-usual socio-economic development scenario (Lenton et al., 2008; Rockström et al., 2017; Schellnhuber et al., 2016). Regarding the recovery timescale it has been shown that human perturbations of the climate system already changed its trajectory on a multi-millennial timescale (Clark et al., 2016; Ganopolski et al., 2016). Therefore a recovery probability per time step p_r close to zero is assumed (Fig. 6.6).

For sufficiently large collapse probabilities (collapse timescale near 30 years and smaller), the climate system is likely to reside in a parameter sweet spot (gray area), where applying an optimization, sustainability or SOS paradigm results in the cautious policy as the advisable way of governing the climate system. However, if the collapse probability per time step is smaller (collapse timescales near 50 years and larger) the situation is different. Here, an SOS and a sustainable paradigm would still yield the cautious policy (Fig. 6.6, cyan), but an optimization paradigm is likely to give the risky policy (Fig. 6.6, red), which at this point is neither sustainable nor safe. One can conclude that in climate policy, economic welfare optimization alone may be neither sustainable nor safe.

Fisheries. For fishery systems, both transition probabilities certainly depend on a variety of factors, e.g. fisher's technical and cultural traits or the dominant fish species in the system, as well as external factors such as climate change influencing habitat condition (Möllmann et al., 2009; Worm et al., 2009). The timescale of a fisheries collapse has been shown to lie within decades (Costello et al., 2008). Roughly consistent with observational and modeled data from the Baltic sea, where the stable regime of high cod biomass lasted approximately from 1970 to 1990 (Möllmann et al., 2009; Österblom et al., 2007), a typical collapse timescale of around 20 years is assumed. Concerning the typical recovery timescale, successful attempts of fish stock recovery lasted for decades (Hutchings and Reynolds, 2004), but is estimated to generally exceed this duration (Caddy and Agnew, 2004). Therefore a larger typical recovery timescale of around 50 years is assumed.

The color gradient in Fig. 6.6 at the fisheries point does not clearly single out a paradigms regime, indicating the dependence on the other parameters at this point. A risky policy might be economically optimal but not sustainable (Fig. 6.6, red). The risky policy eventually leads to the collapse of fish stock (c.f. Costello et al., 2008). At the collapsed and degraded state the conditions for the fishers are not acceptable. Therefore they have to leave the system and cannot wait for the fish's recovery. Yet, further investigation is needed to reduce the uncertainty with respect to the other parameters.

Farming. Last, the case of land degradation by farming is considered. Land degradation and restoration is a complex topic with many influencing factors (Blaikie and Brookfield, 2015). Nevertheless, land degradation by farming has been identified as a tipping element by Kinzig et al. (2006), where the authors discuss the case of the western Australian wheatbelt with a typical collapse timescale of about 100 years. Soil recovery is estimated to take place within 20 to 1000 years (Horrigan et al., 2002), which is roughly consistent with Kinzig et al. (2006), where the duration to reach equilibrium again is estimated with up to 300 years. Therefore, a typical recovery timescale of about 300 years is assumed.

In contrast to climate and fisheries, the transition probabilities for the process of land degradation by farming suggests, that here an optimality paradigm is very likely to yield the risky policy which is neither sustainable nor safe despite considerate parameter uncertainty (red area in Fig. 6.6)

Taken together, it is interesting to see that the climate system in particular may reside at the edge of the parameter regime where economic welfare optimization becomes neither sustainable nor safe (Fig. 6.6). For land degradation by farming, the previous assessment suggests that an optimal policy is likely to yield a non-sustainable and non-safe policy whereas for fisheries the situation is less clear.

6.4 Summary

Overall, this chapter presented a model of a single-agent tipping element environment, designed above the agent-environment interface. This model was used to systematically compare the three governance paradigms of economic optimization, sustainability and safe operating space. The results in this chapter show that in this model, there exists no master paradigm among those three. Policies can be classified by any combination of optimal, sustainable and safe. A master paradigm, in contrast, would guarantee fulfilling requirements imposed by other paradigms. Consequently, the selection of appropriate policy paradigms, especially in more complex settings and models, can be critical for effective environmental governance.

Specifically, it was shown theoretically as well as empirically that economic welfare optimization for managing tipping elements may be neither sustainable nor safe. Theoretically, this is the case, since the volume of the corresponding paradigm combination in parameter space is the largest among those in which the three paradigms actually yield different policies. This suggests the conclusion that the mere structure of a tipping element causes a comparable high chance of obtaining a policy that is neither sustainable nor safe when blindly following an optimization paradigm. Empirically, it was demonstrated that especially in the case of land degradation through farming, optimization may be neither sustainable nor safe. On the other hand, the presented analysis also indicates parameter regimes where economic optimization can safely and sustainably be used. Consider a random tipping element, there is almost a 50% chance that the three paradigms align with each other. Thus, optimization can very well lead to a safe and just operating space, but it is not guaranteed to do so.

The absence of a master paradigm is of special relevance for governing the climate system, since the latter may reside at the edge between parameter regimes where economic welfare optimization becomes neither sustainable nor safe.

Further, simple heuristics were presented, to anticipate when a policy is economic welfare optimal, sustainable and safe. A risky policy may be optimal when the probability of collapse and/or the farsightedness are sufficiently small. It may be sustainable when the probability of a collapse is sufficiently small but the farsightedness is sufficiently large. However, it cannot be safe. A cautious policy may be optimal when the collapse probability and/or the farsightedness are sufficiently large. It is sustainable if its immediate reward exceeds the normatively chosen minimum acceptable reward and it is always safe.

Outlook. Extensions are possible in many directions. Constrained optimization (Altman, 1999) is a straight-forward way to combine the paradigms examined. Policy makers could aim for a policy that delivers the maximum economic welfare and is safe and sustainable, or likewise, least-cost safe target strategies (Ackerman et al., 2009). This is certainly a better approach than relying on economic welfare optimization alone for model-based policy advice. Examples of models for policy advice certainly include integrated assessment models or the use of the maximum sustainable yield

in fisheries management. However, one might not even desire to obtain the welfare optimal safe and sustainable policy in the first place but e.g. the most resilient one, which calls for an operationalization of modern social-ecological resilience concepts (Donges and Barfuss, 2017, P2).

With respect to self-learning agents the results of this chapter call for novel algorithms, which take in a sustainability and safety perspective from the foundations of their design. Such endeavors may have fruitful synergies with the research field of beneficial and safe artificial intelligence (Amodei et al., 2016). A common key challenge here presents the topic of safe exploration, i.e. how a learning agent can safely explore an unknown environment without choosing a truly undesirable action.

The application of the presented model to real-world systems in this chapter is of qualitative, illustrative nature. A more detailed analysis of real world tipping elements in which typical transition probabilities might be estimated from empirical time series could be a way forward to systematize and draw lessons from the multitude of human-environmental tipping elements (Rocha et al., 2014).

Applying this kind of analyses to larger, more complex Markov decision processes would be a way to extend the understanding of the relationships between the paradigms examined. Moreover, it may be desirable to include further policy paradigms into the analyses, e.g. aiming for a large option space of future decision makers (Fleurbaey, 2015; Schellnhuber, 1999). Such analyses may help policy makers in their decisions on how to translate the Sustainable Development Goals and the Paris agreement into concrete policy implementations.

Python code for the reproduction of the reported results plus interactive versions of the figures is available at github: `https://doi.org/10.5281/zenodo.1495578`.

Chapter 7

Conclusion

Ich, Herr Hans Wurst, habe aus meinem Fall einige Lehren gezogen mit knapper Mühe und Not.
Hans Wurst - from Paul Dessau's *Einstein*: Epilogue

7.1 Contributions

This thesis is a physicist's contribution to deepen the theoretical understanding of coupled social-ecological systems, investigating the question what preconditions are required so that collective action towards sustainability can succeed. Thereby, it contributes to the increasing multidisciplinarity of physics (Sinatra et al., 2015) and in particular extends the realm of social (Castellano et al., 2009; Perc et al., 2017) to social-ecological physics.

As the foundational design principle, this thesis proposed the use of the agent-environment interface as known from e.g. Sutton and Barto (1998) for social-ecological systems modeling. As Chapter 2 showed, it is by no means the only design principle upon which social-ecological system models can be build. Yet, doing so has certain advantages. The agent-environment interface presents a unifying perspective across many scientific fields, offering valuable insights for social-ecological systems research. At the same time, it is a comparably simple framework, enabling social-ecological system models to be put into mathematical practice, even with analytical results. The interface offers a clear plug-and-play usage, allowing the straight forward comparison of different rules for the agents' choices and environmental dynamics. This offered valuable insights, as shown in Chapter 4 by comparing different learning dynamics across different environments, or as shown in Chapter 6 by comparing different decision paradigms with each other. As the environment explicitly occurs, the use of the agent-environment interface offered clear guidance in extending social dilemmas in repeated games to social-ecological dilemmas in stochastic games (Chapter 5) and Markov decision processes (Chapter 6).

Before such a social-ecological dilemma could be examined by the means of learning dynamics, learning equations capable of learning in multiple state environments had to be derived. By combining techniques of the statistical physics literature on learning dynamics with established reinforcement learning algorithms from artificial intelligence research, Chapter 4 presented a novel methodological extension to derive a

deterministic limit of so-called temporal difference reinforcement learning algorithms. With respect to the three parameters governing the learning of an agent, Chapter 4 showed that the exploitation level and farsightedness control *where* the learners adapt to in behavior space, weighting current reward, expected future reward and the level of forgetting. The learning rate controls *how fast* the learners adapt along these directions. Demonstrated across multiple example environments, the derived learning equations reveal a variety of dynamical regimes, such as fixed points, periodic orbits and deterministic chaos.

Eventually, these learning equations have been applied to a particular environment, termed the Ecological Public Good, which models a coupled social-ecological dilemma (Chapter 5). Thereby this thesis contributes to a better understanding of social-ecological systems by combining two previously independently studied topics: i) the emergence of cooperation in stochastic games (Hilbe et al., 2018), and ii) so-called collective risk dilemmas (Milinski et al., 2008). As such, the relationship between the farsightedness of the agents and the expected impact in the case of environmental collapse could be examined, with respect to the emergence and stability of cooperation. By a combination of analytical and numerical methods, the theoretical result that cooperation requires a minimum farsightedness could be reproduced (Hilbe et al., 2018). Additionally, each point in behavior space could be associated with a critical farsightedness, above which agents learn to cooperate. Further, the empirical observation that the more severe the expected collapse impact, the more likely the emergence of cooperation (Milinski et al., 2008), could be explained.

With respect to the emergence of cooperation from an initial defective behavior, three qualitatively different parameter regimes could be identified: i) an always cooperate regime, in which risk and damage are so large, that the agents will learn to cooperate, regardless of their farsightedness; ii) a farsightedness regime, where it depends on the farsightedness of the agents, if they learn to cooperate from a full defection initial behavior; and iii) a metaphorical black hole regime, where these agents are not able to learn to cooperate from full defection, even if they have full farsightedness. It was shown that such a black hole regime exits regardless of the collapse risk, if the suffering due to environmental collapse is less than the suffering due to being exploited in the social dilemma. Agents prefer to collectively suffer in environmental collapse, than cooperating in a prosperous environment.

With respect to the stability of cooperation for heterogeneous agents, i.e. when one agent changes its parameters, it was shown through analytical and numerical calculations that an individual learner can keep the cooperation solution stable despite considerate shortsightedness. However, this is only the case if its leverage to collapse the environment and the damages in the case of collapse are large. This means, that a reward optimizing learner who does not believe in likely and severe consequences of a tipping catastrophe will break off the cooperation agreement.

Finally, taking up earlier critiques of optimization approaches for environmental governance, Chapter 6 systematically compares the three decision making paradigms of economic welfare optimization, sustainability and the safe operating space for the governance of a single-agent tipping element environment. It was shown that

optimization alone can lead to safe and sustainable behaviors and policies, but is by no means guaranteed to do so. In fact, there exits no master paradigm among those three, i.e. policies exist in the parameter space of this model which can be classified by any combination of optimal, sustainable and safe. A master paradigm, in contrast, would guarantee fulfilling requirements imposed by other paradigms. Consequently, the selection of appropriate policy paradigms, especially in more complex settings and models, can be critical for effective environmental governance. Further, the absence of a master paradigm is of special relevance for governing the climate system, since the latter may reside at the edge between parameter regimes where economic optimization becomes neither sustainable nor safe.

7.2 Outlook

This thesis demonstrated the usefulness of the agent-environment interface as a design principle for social-ecological system models. Both methodological extensions, as well as further applications to questions regarding the preconditions for sustainability present promising pathways for stimulating future research.

Extension of the agent-environment interface to a multi-layer network perspective. On a technical, methodological level, future work could extend the agent-environment interface as a multi-layer network (Boccaletti et al., 2014). Four different kinds of network layers may encode different kinds of agent-agent interactions (Fig. 7.1).

In a directed action-observation layer, agent i forms a connection to agent j if i is able to observe the current action of agent j. Many models of opinion formation on social networks (e.g. Klemm et al., 2003; San Miguel et al., 2005) can be imagined to happen on this layer if one sees opinions as equivalent to actions. Here, it can be of interest to compare different imitation mechanisms: agents imitate other agents' actions directly vs. agents obtain a positive reward when choosing identical actions (c.f. Banisch and Olbrich, 2018).

In a directed reward-signal-observation layer, agent i is connected to agent j if i is able to observe the reward-signal of j. A model, such as the one presented in Chapter 2, which uses a reward-based imitation scheme utilizes this layer together with the action-observation layer. This reward-signal-observation layer differentiates between a reward signal, that can be in principle observed by other agents, and the true reward an agent perceives to update its behavior. The case when reward and reward signal differ from agent to agent seems of particular interest for future work.

In the directed reward-signal layer, agent i is connected to j if the reward-signal of j depends on i's actions. Typically, evolutionary game theory studies on networks utilize this layer (e.g. Perc et al., 2013; Perc and Szolnoki, 2010; Wang et al., 2015).

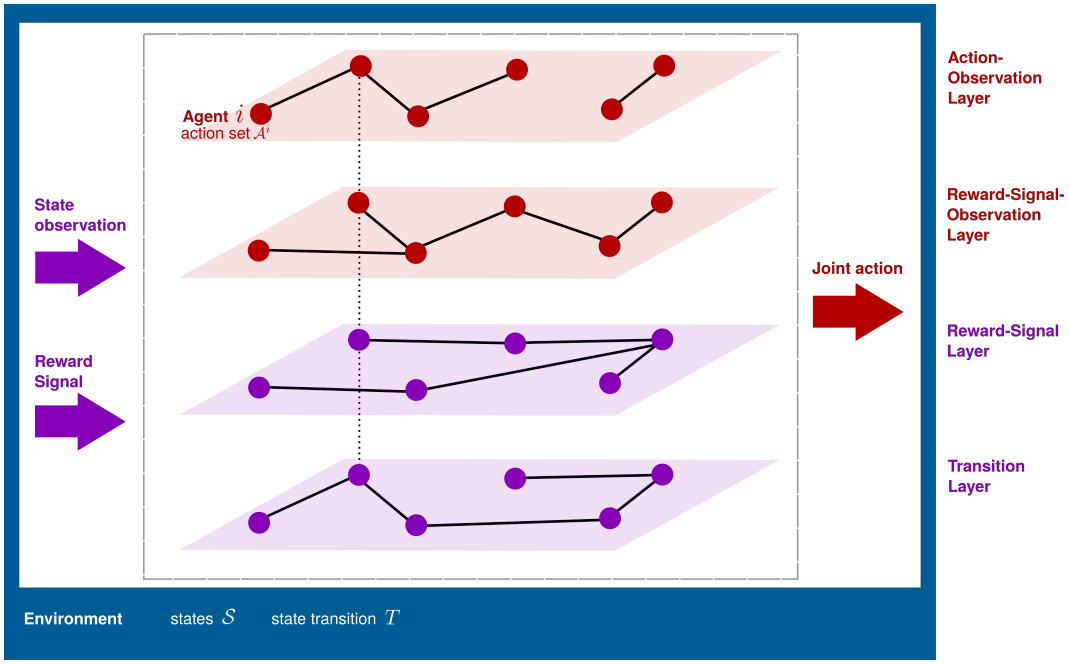

Figure 7.1: Agent-environment interface as a multi-layer network

Finally, in the undirected transition layer agent i and agent j are connected if the state transition probability depends on the joint action of i and j. Conceptually, such a transition layer is similar to a Markov random field (Lauritzen, 1996), which is an interesting connection to be explored.

Thus, perceiving the agent-environment interface as a multi-layer network would have the advantage to present an even more unifying perspective, additionally incorporating scientific fields, such as models of opinion formation (e.g. Klemm et al., 2003), evolutionary dynamics on networks (e.g. Perc and Szolnoki, 2010), and even graphical models (e.g. Lauritzen, 1996). As such, it may be a starting point to reconcile models, similar to the one presented in Chapter 2 with the models of the remainder of this thesis within a common framework.

Further, within this framework future work could extend the dynamical systems description presented in Chapter 4 to partial observability of the Markov states of the environment (Oliehoek, 2012; Spaan, 2012), behavior profiles with history, and other-regarding agent (i.e. joint-action) learners (c.f. Busoniu et al., 2008 for an overview of other-regarding agent learning algorithms). Also agents that learn an explicit model of the environment (Hester and Stone, 2012) might turn out to be useful for social-ecological systems research. Further, the combination of individual and social learning (Bandura, 1977; Banisch and Olbrich, 2018; Smolla et al., 2015) seems promising.

Application to questions related to collective action for sustainability. Within the perspective of the agent-environment interface as a multi-layer network, relevant research questions can be asked. For example, with respect to self-learning agents novel algorithms are of great interest which take in a sustainability and safety perspective from the foundations of their design. Endeavors to find such algorithms may have fruitful synergies with the research field of beneficial and safe artificial intelligence (Amodei et al., 2016). A common key challenge here is the topic of safe exploration, i.e. how a learning agent can safely explore an unknown environment without choosing a truly undesirable action. Moving forward with operationalizing modern facets of social-ecological resilience, as proposed by Donges and Barfuss (2017, P2), may also contribute to finding safe and sustainable learning algorithms.

Further of interest may be to study how learning agents compare when applied in different environments, each presenting a challenge for sustainability. For example, such environments may represent the harvesting of common-pool renewable resources (Lindkvist and Norberg, 2014; Schill et al., 2015) or the prevention of dangerous climate change (Barrett and Dannenberg, 2012; Milinski et al., 2008). Such studies, in which these challenges are examined individually, could lead to further, more complicated environments, in which such individual challenges are subsequently combined. Here, examining effects of heterogeneous agents and inequality (Vasconcelos et al., 2014) remain of great interest.

The studies presented in this thesis which used the agent-environment interface had a strong behavioral focus. Social institutions and norms (Nyborg et al., 2016) have not yet been purposely included, albeit their importance for collective action towards sustainability. Future work could explore the mechanisms how social norms emergence from the agent's collective behavior.

Last, an important descriptive question concerns the modeling of human behavior in social-ecological system models (Schlüter et al., 2017). Here, two questions may turn out to be productive: i) What characteristics does a model of human behavior have to fulfill in order to be appropriate for the use in social-ecological system models? ii) Which model fulfills these characteristics best? With respect to the first question, social science theories of human decision making in combination with behavioral experiments could lead to a computational test suite. Towards answering the second question, computational models of human behavior could be compared within this test suite, receiving each a score measuring how well they performed across multiple different environments. Subsequently extending this test suit may thus lead to increasingly better models of human behavior for social-ecological system models.

All together, such endeavors could lead to a deeper theoretical understanding of social-ecological systems to be put into practical use to find and influence the preconditions for successful collective action towards ecologically safe and socially just sustainability.

Bibliography

Ackerman, F., S. J. DeCanio, R. B. Howarth, and K. Sheeran (2009). "Limitations of integrated assessment models of climate change". In: *Climatic Change* 95.3, pp. 297–315. DOI: 10.1007/s10584-009-9570-x.

Akiyama, E. and K. Kaneko (2000). "Dynamical systems game theory and dynamics of games". In: *Physica D: Nonlinear Phenomena* 147.3-4, pp. 221–258. DOI: 10.1016/S0167-2789(00)00157-3.

– (2002). "Dynamical systems game theory II: A new approach to the problem of the social dilemma". In: *Physica D: Nonlinear Phenomena* 167.1-2, pp. 36–71. DOI: 10.1016/S0167-2789(02)00402-5.

Aloric, A., P. Sollich, P. McBurney, and T. Galla (2016). "Emergence of Cooperative Long-Term Market Loyalty in Double Auction Markets". In: *PloS ONE* 11.4, e0154606. DOI: https://doi.org/10.1371/journal.pone.0154606.

Altman, E. (1999). *Constrained Markov decision processes*. Chapman & Hall/CRC.

Amodei, D., C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané (2016). "Concrete problems in AI safety". In: *ArXiv preprint* arXiv:1606.06565.

Anderies, J. M., A. A. Rodriguez, M. A. Janssen, and O. Cifdaloz (2007). "Panaceas, uncertainty, and the robust control framework in sustainability science". In: *Proceedings of the National Academy of Sciences* 104.39, pp. 15194–15199. DOI: 10.1073/pnas.0702655104.

Arrow, K. J., P. Dasgupta, L. H. Goulder, K. J. Mumford, and K. Oleson (2012). "Sustainability and the measurement of wealth". In: *Environment and Development Economics* 17.03, pp. 317–353. DOI: 10.1017/S1355770X12000137.

Arthur, W. B. (1993). "On designing economic agents that behave like human agents". In: *Journal of Evolutionary Economics* 3.1, pp. 1–22. DOI: 10.1007/bf01199986.

– (1999). "Complexity and the economy". In: *Science* 284.5411, pp. 107–109. DOI: 10.1126/science.284.5411.107.

Aubin, J.-P., A. M. Bayen, and P. Saint-Pierre (2011). *Viability Theory*. Springer Berlin Heidelberg. DOI: 10.1007/978-3-642-16684-6.

Auer, S, J Heitzig, U Kornek, E Schöll, and J Kurths (2015). "The Dynamics of Coalition Formation on Complex Networks". In: *Scientific Reports* 5, pp. 1–7. DOI: 10.1038/srep13386.

Axelrod, R and W. Hamilton (1981). "The evolution of cooperation". In: *Science* 211.4489, pp. 1390–1396. DOI: 10.1126/science.7466396.

Bahar, D., R. Hausmann, and C. A. Hidalgo (2014). "Neighbors and the evolution of the comparative advantage of nations: Evidence of international knowledge diffusion?" In: *Journal of International Economics* 92.1, pp. 111–123. DOI: 10.1016/j.jinteco.2013.11.001.

Bandura, A. (1977). *Social learning Theory*. Englewood Cliffs, NJ: Prentice Hall.

Banisch, S. and E. Olbrich (2018). "Opinion polarization by learning from social feedback". In: *The Journal of Mathematical Sociology* 0.0, pp. 1–28. DOI: 10.1080/0022250X.2018.1517761.

Barfuss, W., J. F. Donges, M. Wiedermann, and W. Lucht (2017, P1). "Sustainable use of renewable resources in a stylized social–ecological network model under heterogeneous resource distribution". In: *Earth System Dynamics* 8.2, pp. 255–264. DOI: 10.5194/esd-8-255-2017.

Barfuss, W., J. F. Donges, S. J. Lade, and J. Kurths (2018, P4). "When optimization for governing human-environment tipping elements is neither sustainable nor safe". In: *Nature communications* 9.1, p. 2354. DOI: 10.1038/s41467-018-04738-z.

Barfuss, W., J. F. Donges, and J. Kurths (2019, P7). "Deterministic limit of temporal difference reinforcement learning for stochastic games". In: *Physical Review E* 99.4, p. 043305.

Barfuss, W., J. F. Donges, J. Kurths, and S. Levin (in prep., P8). "On the emergence and stability of cooperation in the ecological public good". In: *preparation*.

Barrett, S. and A. Dannenberg (2012). "Climate negotiations under scientific uncertainty". In: *Proceedings of the National Academy of Sciences* 109.43, pp. 17372–17376. DOI: 10.1073/pnas.1208417109.

Barrett, S. and A. Dannenberg (2013). "Sensitivity of collective action to uncertainty about climate tipping points". In: *Nature Climate Change* 4.1, pp. 36–39. DOI: 10.1038/nclimate2059.

Bechtel, M. M., T. Bernauer, and R. Meyer (2012). "The green side of protectionism: Environmental concerns and three facets of trade policy preferences". In: *Review of International Political Economy* 19.5, pp. 837–866. DOI: 10.1080/09692290.2011.611054.

Bellman, R. (1957). "A Markovian Decision Process". In: *Indiana University Mathematics Journal* 6.4, pp. 679–684. DOI: 10.1512/iumj.1957.6.56038.

Bentley, R. A., E. J. Maddison, P. H. Ranner, J. Bissell, C. C. S. Caiado, P. Bhatanacharoen, T. Clark, M. Botha, F. Akinbami, M. Hollow, R. Michie, B. Huntley, S. E. Curtis, and P. Garnett (2014). "Social tipping points and Earth systems dynamics". In: *Frontiers in Environmental Science* 2, p. 35. DOI: 10.3389/fenvs.2014.00035.

Berkes, F. and C. Folke (1998). *Linking social and ecological systems: management practices and social mechanisms for building resilience*. Cambridge University Press.

Bladon, A. J. and T. Galla (2011). "Learning dynamics in public goods games". In: *Physical Review E* 84.4. DOI: 10.1103/physreve.84.041132.

Blaikie, P. and H. Brookfield (2015). *Land Degradation and Society*. Routledge. DOI: 10.4324/9781315685366.

Bloembergen, D., K. Tuyls, D. Hennes, and M. Kaisers (2015). "Evolutionary Dynamics of Multi-Agent Learning: A Survey". In: *Journal of Artificial Intelligence Research* 53, pp. 659–697. DOI: 10.1613/jair.4818.

Boccaletti, S., G. Bianconi, R. Criado, C. del Genio, J. Gómez-Gardeñes, M. Romance, I. Sendiña-Nadal, Z. Wang, and M. Zanin (2014). "The structure and dynamics of multilayer networks". In: *Physics Reports* 544.1, pp. 1–122. DOI: 10.1016/j.physrep.2014.07.001.

Bodin, Ö. and M. Tengö (2012). "Disentangling intangible social–ecological systems". In: *Global Environmental Change* 22.2, pp. 430–439. DOI: 10.1016/j.gloenvcha.2012.01.005.

Börgers, T. and R. Sarin (1997). "Learning Through Reinforcement and Replicator Dynamics". In: *Journal of Economic Theory* 77.1, pp. 1–14. DOI: 10.1006/jeth.1997.2319.

Brander, J. A. and M. S. Taylor (1998). "The Simple Economics of Easter Island: A Ricardo-Malthus Model of Renewable Resource Use". In: *The American Economic Review* 88.1, pp. 119–138.

Branke, J., K. Deb, K. Miettinen, and R. Słowiński (2008). *Multiobjective Optimization*. Springer Berlin Heidelberg. DOI: 10.1007/978-3-540-88908-3.

Busoniu, L., R. Babuska, and B. D. Schutter (2008). "A Comprehensive Survey of Multiagent Reinforcement Learning". In: *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 38.2, pp. 156–172. DOI: 10.1109/tsmcc.2007.913919.

Caddy, J. and D. Agnew (2004). "An overview of recent global experience with recovery plans for depleted marine resources and suggested guidelines for recovery planning". In: *Reviews in Fish Biology and Fisheries* 14.1, pp. 43–112. DOI: 10.1007/s11160-004-3770-2.

Cai, Y., T. M. Lenton, and T. S. Lontzek (2016). "Risk of multiple interacting tipping points should encourage rapid CO2 emission reduction". In: *Nature Climate Change* 6.5, pp. 520–525. DOI: 10.1038/nclimate2964.

Camerer, C. and T. H. Ho (1999). "Experience-weighted Attraction Learning in Normal Form Games". In: *Econometrica* 67.4, pp. 827–874. DOI: 10.1111/1468-0262.00054.

Camerer, C. F. (2003). *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton University Press.

Camerer, C. F., G. Loewenstein, and M. Rabin (2004). *Advances in Behavioral Economics*. Princeton University Press.

Capraro, V. and M. Perc (2018). "Grand Challenges in Social Physics: In Pursuit of Moral Behavior". In: *Frontiers in Physics* 6. DOI: 10.3389/fphy.2018.00107.

Carpenter, S. R., W. A. Brock, C. Folke, E. H. van Nes, and M. Scheffer (2015). "Allowing variance may enlarge the safe operating space for exploited ecosystems". In: *Proceedings of the National Academy of Sciences* 112.46, pp. 14384–14389. DOI: 10.1073/pnas.1511804112.

Castellano, C., S. Fortunato, and V. Loreto (2009). "Statistical physics of social dynamics". In: *Reviews of Modern Physics* 81.2, pp. 591–646. DOI: 10.1103/revmodphys.81.591.

Centola, D., J. C. Gonzalez-Avella, V. M. Eguiluz, M. San Miguel, D. Centola, J. C. González-Avella, V. M. Eguíluz, and M. S. Miguel (2007). "Homophily, Cultural Drift, and the Co-Evolution of Cultural Groups". In: *Journal of Conflict Resolution* 51.6, pp. 905–929. DOI: 10.1177/0022002707307632.

Chadès, I., S. Nicol, T. M. Rout, M. Péron, Y. Dujardin, J.-B. Pichancourt, A. Hastings, and C. E. Hauser (2016). "Optimization methods to solve adaptive management problems". In: *Theoretical Ecology* 10.1, pp. 1–20. DOI: 10.1007/s12080-016-0313-0.

Chung, N. N., L. Y. Chew, and C. H. Lai (2013). "Influence of network structure on cooperative dynamics in coupled socio-ecological systems". In: *EPL (Europhysics Letters)* 104.2, p. 28003. DOI: 10.1209/0295-5075/104/28003.

Clark, P. U., J. D. Shakun, S. A. Marcott, A. C. Mix, M. Eby, S. Kulp, A. Levermann, G. A. Milne, P. L. Pfister, B. D. Santer, D. P. Schrag, S. Solomon, T. F. Stocker, B. H. Strauss, A. J. Weaver, R. Winkelmann, D. Archer, E. Bard, A. Goldner, K. Lambeck, R. T. Pierrehumbert, and G.-K. Plattner (2016). "Consequences of twenty-first-century policy for multi-millennial climate and sea-level change". In: *Nature Climate Change* 6.4, pp. 360–369. DOI: 10.1038/nclimate2923.

Claussen, M., L. Mysak, A. Weaver, M. Crucifix, T. Fichefet, M.-F. Loutre, S. Weber, J. Alcamo, V. Alexeev, A. Berger, R. Calov, A. Ganopolski, H. Goosse, G. Lohmann, F. Lunkeit, I. Mokhov, V. Petoukhov, P. Stone, and Z. Wang (2002). "Earth system models of intermediate complexity: closing the gap in the spectrum of climate system models". In: *Climate Dynamics* 18.7, pp. 579–586. DOI: 10.1007/s00382-001-0200-1.

COP (2015). *Conference of the Parties - Adoption of the Paris Agreement*. Adoption. United Nations.

Costello, C., S. D. Gaines, and J. Lynham (2008). "Can Catch Shares Prevent Fisheries Collapse?" In: *Science* 321.5896, pp. 1678–1681. DOI: 10.1126/science.1159478.

Cross, J. G. (1973). "A Stochastic Learning Model of Economic Behavior". In: *The Quarterly Journal of Economics* 87.2, p. 239. DOI: 10.2307/1882186.

Crutzen, P. J. (2002). "Geology of mankind". In: *Nature* 415.6867, p. 23. DOI: 10.1038/415023a.

Dannenberg, A., A. Löschel, G. Paolacci, C. Reif, and A. Tavoni (2014). "On the Provision of Public Goods with Probabilistic and Ambiguous Thresholds". In: *Environmental and Resource Economics* 61.3, pp. 365–383. DOI: 10.1007/s10640-014-9796-6.

Dasgupta, P. and K.-G. Mäler (2004). *The Economics of Non-Convex Ecosystems*. Springer Netherlands. DOI: 10.1007/1-4020-2515-7.

Dawes, R. M. (1980). "Social Dilemmas". In: *Annual Review of Psychology* 31.1, pp. 169–193. DOI: 10.1146/annurev.ps.31.020180.001125.

De Lara, M. and L. Doyen (2008). *Sustainable Management of Natural Resources*. Springer-Verlag Berlin Heidelberg. DOI: 10.1007/978-3-540-79074-7.

Dearing, J. A., R. Wang, K. Zhang, J. G. Dyke, H. Haberl, M. S. Hossain, P. G. Langdon, T. M. Lenton, K. Raworth, S. Brown, J. Carstensen, M. J. Cole, S. E. Cornell, T. P. Dawson, C. P. Doncaster, F. Eigenbrod, M. Flörke, E. Jeffers, A. W. Mackay, B. Nykvist, and G. M. Poppy (2014). "Safe and just operating spaces for regional social-ecological systems". In: *Global Environmental Change* 28, pp. 227–238. DOI: 10.1016/j.gloenvcha.2014.06.012.

Derissen, S., M. F. Quaas, and S. Baumgärtner (2011). "The relationship between resilience and sustainability of ecological-economic systems". In: *Ecological Economics* 70.6, pp. 1121–1128. DOI: 10.1016/j.ecolecon.2011.01.003.

Donges, J. F. and W. Barfuss (2017, P2). "From Math to Metaphors and Back Again: Social-Ecological Resilience from a Multi-Agent-Environment Perspective". In: *GAIA - Ecological Perspectives for Science and Society* 26.1, pp. 182–190. DOI: 10.14512/gaia.26.s1.5.

Donges, J. F., R. Winkelmann, W. Lucht, S. E. Cornell, J. G. Dyke, J. Rockström, J. Heitzig, and H. J. Schellnhuber (2017). "Closing the loop: Reconnecting human dynamics to Earth System science". In: *The Anthropocene Review* 4.2, pp. 151–157. DOI: 10.1177/2053019617725537.

Donges, J. F., J. Heitzig, W. Barfuss, J. A. Kassel, T. Kittel, J. J. Kolb, T. Kolster, F. Müller-Hansen, I. M. Otto, M. Wiedermann, et al. (2018, P5). "Earth system modeling with complex dynamic human societies: the copan: CORE World-Earth modeling framework". In: *Earth System Dynamics Discussions* 2018, pp. 1–27. DOI: 10.5194/esd-2017-126.

Donges, J. F., W. Lucht, J. Heitzig, W. Barfuss, S. E. Cornell, S. J. Lade, and M. Schlüter (2018, P6). "Taxonomies for structuring models for World-Earth system

analysis of the Anthropocene: subsystems, their interactions and social-ecological feedback loops". In: *Earth System Dynamics Discussions* 2018, pp. 1–30. DOI: 10.5194/esd-2018-27.

Dunlap, R. E., A. M. McCright, and J. H. Yarosh (2016). "The Political Divide on Climate Change: Partisan Polarization Widens in the U.S." In: *Environment: Science and Policy for Sustainable Development* 58.5, pp. 4–23. DOI: 10.1080/00139157.2016.1208995.

Ehrgott, M. (2000). *Multicriteria Optimization*. Springer Berlin Heidelberg. DOI: 10.1007/978-3-662-22199-0.

Erdös, P. and A Rényi (1960). "On the evolution of random graphs". In: *Publ. Math. Inst. Hungar. Acad. Sci* 5, pp. 17–61.

Erev, I. and A. E. Roth (1998). "Predicting How People Play Games: Reinforcement Learning in Experimental Games with Unique, Mixed Strategy Equilibria". In: *The American Economic Review* 88.4, pp. 848–881.

Ewing, B., S. Goldfinger, M. Wackernagel, M. Stechbart, S. M. Rizk, A. Reed, and J. Kitzes (2008). *The Ecological Footprint Atlas 2008*. Oakland: Global Footprint Network.

Farmer, J. D., C. Hepburn, P. Mealy, and A. Teytelboym (2015). "A Third Wave in the Economics of Climate Change". In: *Environmental and Resource Economics* 62.2, pp. 329–357. DOI: 10.1007/s10640-015-9965-2.

Fischer, J., G. D. Peterson, T. A. Gardner, L. J. Gordon, I. Fazey, T. Elmqvist, A. Felton, C. Folke, and S. Dovers (2009). "Integrating resilience thinking and optimisation for conservation". In: *Trends in Ecology & Evolution* 24.10, pp. 549–554. DOI: 10.1016/j.tree.2009.03.020.

Fleurbaey, M. (2015). "On sustainability and social welfare". In: *Journal of Environmental Economics and Management* 71, pp. 34–53. DOI: 10.1016/j.jeem.2015.02.005.

Folke, C., S. R. Carpenter, B. Walker, M. Scheffer, T. Chapin, and J. Rockström (2010). "Resilience Thinking: Integrating Resilience, Adaptability and Transformability". In: *Ecology and Society* 15.4. DOI: 10.5751/es-03610-150420.

Fudenberg, D. and D. K. Levine (1998). *The Theory of Learning in Games*. MIT Press.

Galla, T. (2009). "Intrinsic Noise in Game Dynamical Learning". In: *Physical Review Letters* 103.19. DOI: 10.1103/physrevlett.103.198702.

– (2011). "Cycles of cooperation and defection in imperfect learning". In: *Journal of Statistical Mechanics: Theory and Experiment* 2011.08, P08007. DOI: 10.1088/1742-5468/2011/08/p08007.

Galla, T. and J. D. Farmer (2013). "Complex dynamics in learning complicated games". In: *Proceedings of the National Academy of Sciences* 110.4, pp. 1232–1236. DOI: 10.1073/pnas.1109672110.

Ganopolski, A., R. Winkelmann, and H. J. Schellnhuber (2016). "Critical insolation–CO2 relation for diagnosing past and future glacial inception". In: *Nature* 529.7585, pp. 200–203. DOI: 10.1038/nature16494.

Gerlagh, R. (2017). "Generous Sustainability". In: *Ecological Economics* 136, pp. 94–100. DOI: 10.1016/j.ecolecon.2017.02.012.

Gintis, H. (2014). *The Bounds of Reason*. Princeton University Press. DOI: 10.1515/9781400851348.

Greco, S., M. Ehrgott, and J. R. Figueira (2016). *Multiple Criteria Decision Analysis*. Springer New York. DOI: 10.1007/978-1-4939-3094-4.

Griggs, D., M. Stafford-Smith, O. Gaffney, J. Rockström, M. C. Öhman, P. Shyamsundar, W. Steffen, G. Glaser, N. Kanie, and I. Noble (2013). "Sustainable development goals for people and planet". In: *Nature* 495.7441, pp. 305–307. DOI: 10.1038/495305a.

Gross, T. and B. Blasius (2008). "Adaptive coevolutionary networks: a review". In: *Journal of The Royal Society Interface* 5.20, pp. 259–271. DOI: 10.1098/rsif.2007.1229.

Hardin, G. (1968). "The Tragedy of the Commons". In: *Science* 162.3859, pp. 1243–1248. DOI: 10.1126/science.162.3859.1243.

Hassabis, D., D. Kumaran, C. Summerfield, and M. Botvinick (2017). "Neuroscience-Inspired Artificial Intelligence". In: *Neuron* 95.2, pp. 245–258. DOI: 10.1016/j.neuron.2017.06.011.

Heitzig, J., T. Kittel, J. F. Donges, and N. Molkenthin (2016). "Topology of sustainable management of dynamical systems with desirable states: from defining planetary boundaries to safe operating spaces in the Earth system". In: *Earth System Dynamics* 7.1, pp. 21–50. DOI: 10.5194/esd-7-21-2016.

Heitzig, J., W. Barfuss, and J. F. Donges (2018, P3). "A Thought Experiment on Sustainable Management of the Earth System". In: *Sustainability* 10.6. DOI: 10.3390/su10061947.

Hennes, D., K. Tuyls, and M. Rauterberg (2009). "State-coupled replicator dynamics". In: *Proceedings of the 8th International Conference on Autonomous Agents and Multiagent Systems*. AAMAS 2009, pp. 789–796.

Hennes, D., M. Kaisers, and K. Tuyls (2010). "RESQ-learning in stochastic games". In: *Proceedings of the Adaptive and Learning Agents Workshop*. ALA 2010, pp. 8–15.

Hester, T. and P. Stone (2012). "Learning and Using Models". In: *Reinforcement Learning*. Springer Berlin Heidelberg, pp. 111–141. DOI: 10.1007/978-3-642-27645-3_4.

Hilbe, C., Š. Šimsa, K. Chatterjee, and M. A. Nowak (2018). "Evolution of cooperation in stochastic games". In: *Nature* 559.7713, pp. 246–249. DOI: 10.1038/s41586-018-0277-x.

Hinkel, J., P. W. Bots, and M. Schlüter (2014). "Enhancing the Ostrom social-ecological system framework through formalization". In: *Ecology and Society* 19.3. DOI: 10.5751/es-06475-190351.

Holme, P. and M. E. J. Newman (2006). "Nonequilibrium phase transition in the coevolution of networks and opinions". In: *Physical Review E* 74.5. DOI: 10.1103/physreve.74.056108.

Horrigan, L., R. S. Lawrence, and P. Walker (2002). "How Sustainable Agriculture Can Address the Environmental and Human Health Harms of Industrial Agriculture". In: *Environmental Health Perspectives* 110.5, pp. 445–456. DOI: 10.1289/ehp.02110445.

Hutchings, J. A. and J. D. Reynolds (2004). "Marine Fish Population Collapses: Consequences for Recovery and Extinction Risk". In: *BioScience* 54.4, pp. 297–309. DOI: 10.1641/0006-3568(2004)054[0297:MFPCCF]2.0.CO;2.

Irwin, E. G., S. Gopalakrishnan, and A. Randall (2016). "Welfare, Wealth, and Sustainability". In: *Annual Review of Resource Economics* 8.1, pp. 77–98. DOI: 10.1146/annurev-resource-100815-095351.

Janssen, M. A., Ö. Bodin, J. M. Anderies, T. Elmqvist, H. Ernstson, R. R. McAllister, P. Olsson, and P. Ryan (2006). "Toward a Network Perspective of the Study of Resilience in Social-Ecological Systems". In: *Ecology and Society* 11.1. DOI: 10.5751/es-01462-110115.

Kaisers, M. and K. Tuyls (2010). "Frequency adjusted multi-agent Q-learning". In: *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: Volume 1*. AAMAS 2010, pp. 309–315.

Keeling, M. J. (2000). "Multiplicative Moments and Measures of Persistence in Ecology". In: *Journal of Theoretical Biology* 205.2, pp. 269–281. DOI: 10.1006/jtbi.2000.2066.

Kinzig, A. P., P. Ryan, M. Etienne, H. Allison, T. Elmqvist, and B. H. Walker (2006). "Resilience and Regime Shifts: Assessing Cascading Effects". In: *Ecology and Society* 11.1. DOI: 10.5751/es-01678-110120.

Kirk, D. E. (2012). *Optimal control theory: an introduction*. Dova Publications, Inc. Mineola, New York.

Klemm, K., V. M. Eguíluz, R. Toral, and M. San Miguel (2003). "Global culture: A noise-induced transition in finite systems". In: *Physical Review E* 67.4. DOI: 10.1103/physreve.67.045101.

Lade, S. J., A. Tavoni, S. A. Levin, and M. Schlüter (2013). "Regime shifts in a social-ecological system". In: *Theoretical Ecology* 6.3, pp. 359–372. DOI: 10.1007/s12080-013-0187-3.

Lade, S. J., Ö. Bodin, J. F. Donges, E. E. Kautsky, D. Galafassi, P. Olsson, and M. Schlüter (2017). "Modelling social-ecological transformations: an adaptive network proposal". In: *ArXiv preprint* arXiv:1704.06135.

Lange, S., T. Gabel, and M. Riedmiller (2012). "Batch Reinforcement Learning". In: *Reinforcement Learning*. Springer Berlin Heidelberg, pp. 45–73. DOI: 10.1007/978-3-642-27645-3_2.

Lauritzen, S. L. (1996). *Graphical models*. Vol. 17. Clarendon Press, Oxford.

Leibo, J. Z., V. Zambaldi, M. Lanctot, J. Marecki, and T. Graepel (2017). "Multi-agent Reinforcement Learning in Sequential Social Dilemmas". In: *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*. AAMAS 2017, pp. 464–473.

Lenton, T. M., H. Held, E. Kriegler, J. W. Hall, W. Lucht, S. Rahmstorf, and H. J. Schellnhuber (2008). "Tipping elements in the Earth's climate system". In: *Proceedings of the National Academy of Sciences* 105.6, pp. 1786–1793. DOI: 10.1073/pnas.0705414105.

Leslie, H. M., X. Basurto, M. Nenadovic, L. Sievanen, K. C. Cavanaugh, J. J. Cota-Nieto, B. E. Erisman, E. Finkbeiner, G. Hinojosa-Arango, M. Moreno-Báez, S. Nagavarapu, S. M. W. Reddy, A. Sánchez-Rodríguez, K. Siegel, J. J. Ulibarria-Valenzuela, A. H. Weaver, and O. Aburto-Oropeza (2015). "Operationalizing the social-ecological systems framework to assess sustainability". In: *Proceedings of the National Academy of Sciences* 112.19, pp. 5979–5984. DOI: 10.1073/pnas.1414640112.

Levin, S. (2013). "The mathematics of sustainability". In: *Notices of the American Mathematical Society* 60.04, p. 1. DOI: 10.1090/noti982.

Levin, S., T. Xepapadeas, A.-S. Crépin, J. Norberg, A. de Zeeuw, C. Folke, T. Hughes, K. Arrow, S. Barrett, G. Daily, P. Ehrlich, N. Kautsky, K.-G. Mäler, S. Polasky, M. Troell, J. R. Vincent, and B. Walker (2012). "Social-ecological systems as complex adaptive systems: modeling and policy implications". In: *Environment and Development Economics* 18.02, pp. 111–132. DOI: 10.1017/s1355770x12000460.

Lindahl, T., A.-S. Crépin, and C. Schill (2016). "Potential Disasters can Turn the Tragedy into Success". In: *Environmental and Resource Economics* 65.3, pp. 657–676. DOI: 10.1007/s10640-016-0043-1.

Lindkvist, E. and J. Norberg (2014). "Modeling experiential learning: The challenges posed by threshold dynamics for sustainable renewable resource management". In: *Ecological Economics* 104, pp. 107–118. DOI: 10.1016/j.ecolecon.2014.04.018.

Lindkvist, E., Ö. Ekeberg, and J. Norberg (2017). "Strategies for sustainable management of renewable resources during environmental change". In: *Proceedings of the Royal Society B: Biological Sciences* 284.1850, p. 20162762. DOI: 10.1098/rspb.2016.2762.

Lontzek, T. S., Y. Cai, K. L. Judd, and T. M. Lenton (2015). "Stochastic integrated assessment of climate tipping points indicates the need for strict climate policy". In: *Nature Climate Change* 5.5, pp. 441–444. DOI: 10.1038/nclimate2570.

Macy, M. W. and A. Flache (2002). "Learning dynamics in social dilemmas". In: *Proceedings of the National Academy of Sciences* 99.Supplement 3, pp. 7229–7236. DOI: 10.1073/pnas.092080099.

Mäler, K.-G. and C.-Z. Li (2010). "Measuring sustainability under regime shift uncertainty: a resilience pricing approach". In: *Environment and Development Economics* 15.06, pp. 707–719. DOI: 10.1017/s1355770x10000318.

Malthus, T. R. (1798). *An essay on the principle of population, as it affects the future improvement of society. With remarks on the speculations of Mr. Godwin, M. Condorcet and other writers.* London, J. Johnson.

Marsili, M., D. Challet, and R. Zecchina (2000). "Exact solution of a modified El Farol's bar problem: Efficiency and the role of market impact". In: *Physica A: Statistical Mechanics and its Applications* 280.3-4, pp. 522–553. DOI: 10.1016/s0378-4371(99)00610-x.

Martinet, V. and L. Doyen (2007). "Sustainability of an economy with an exhaustible resource: A viable control approach". In: *Resource and Energy Economics* 29.1, pp. 17–39. DOI: 10.1016/j.reseneeco.2006.03.003.

McPherson, M., L. Smith-Lovin, and J. M. Cook (2001). "Birds of a Feather: Homophily in Social Networks". In: *Annual Review of Sociology* 27.1, pp. 415–444. DOI: 10.1146/annurev.soc.27.1.415.

Meadows, D. H., D. L. Meadows, J. Randers, and W. W. Behrens III (1972). *The limits to growth.* Universe Books New York. DOI: 10.1349/ddlp.1.

Meurer, A., C. P. Smith, M. Paprocki, O. Čertík, S. B. Kirpichev, M. Rocklin, A. Kumar, S. Ivanov, J. K. Moore, S. Singh, T. Rathnayake, S. Vig, B. E. Granger, R. P. Muller, F. Bonazzi, H. Gupta, S. Vats, F. Johansson, F. Pedregosa, M. J. Curry, A. R. Terrel, v. Roučka, A. Saboo, I. Fernando, S. Kulal, R. Cimrman, and A. Scopatz (2017). "SymPy: symbolic computing in Python". In: *PeerJ Computer Science* 3, e103. DOI: 10.7717/peerj-cs.103.

Milinski, M., R. D. Sommerfeld, H.-J. Krambeck, F. A. Reed, and J. Marotzke (2008). "The collective-risk social dilemma and the prevention of simulated dangerous climate change". In: *Proceedings of the National Academy of Sciences* 105.7, pp. 2291–2294. DOI: 10.1073/pnas.0709546105.

Mnih, V., K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis (2015). "Human-level control through deep reinforcement learning". In: *Nature* 518.7540, pp. 529–533. DOI: 10.1038/nature14236.

Möllmann, C., R. Diekmann, B. Müller-Karulis, G. Kornilovs, M. Plikshs, and P. Axe (2009). "Reorganization of a large marine ecosystem due to atmospheric and anthropogenic pressure: a discontinuous regime shift in the Central Baltic Sea". In: *Global Change Biology* 15.6, pp. 1377–1393. DOI: 10.1111/j.1365-2486.2008.01814.x.

Neumayer, E. (2003). *Weak versus strong sustainability: exploring the limits of two opposing paradigms*. Edward Elgar, Cheltenham, UK.

Newig, J., D. Günther, C. Pahl-Wostl, et al. (2010). "Synapses in the Network: Learning in Governance Networks in the Context of Environmental Management". In: *Ecology and Society* 15.4. DOI: 10.5751/es-03713-150424.

Neyman, A. and S. Sorin (2003). *Stochastic Games and Applications*. Springer Netherlands. DOI: 10.1007/978-94-010-0189-2.

Nordhaus, W. D. (2007). "A Review of theStern Review on the Economics of Climate Change". In: *Journal of Economic Literature* 45.3, pp. 686–702. DOI: 10.1257/jel.45.3.686.

Nowak, M. A. (2006). "Five Rules for the Evolution of Cooperation". In: *Science* 314.5805, pp. 1560–1563. DOI: 10.1126/science.1133755.

Nyborg, K., J. M. Anderies, A. Dannenberg, T. Lindahl, C. Schill, M. Schlüter, W. N. Adger, K. J. Arrow, S. Barrett, S. Carpenter, F. S. Chapin, A.-S. Crépin, G. Daily, P. Ehrlich, C. Folke, W. Jager, N. Kautsky, S. A. Levin, O. J. Madsen, S. Polasky, M. Scheffer, B. Walker, E. U. Weber, J. Wilen, A. Xepapadeas, and A. de Zeeuw (2016). "Social norms as solutions". In: *Science* 354.6308, pp. 42–43. DOI: 10.1126/science.aaf8317.

Oliehoek, F. A. (2012). "Decentralized POMDPs". In: *Adaptation, Learning, and Optimization*. Springer Berlin Heidelberg, pp. 471–503. DOI: 10.1007/978-3-642-27645-3_15.

Österblom, H., S. Hansson, U. Larsson, O. Hjerne, F. Wulff, R. Elmgren, and C. Folke (2007). "Human-induced Trophic Cascades and Ecological Regime Shifts in the Baltic Sea". In: *Ecosystems* 10.6, pp. 877–889. DOI: 10.1007/s10021-007-9069-0.

Ostrom, E. (2007). "A diagnostic approach for going beyond panaceas". In: *Proceedings of the National Academy of Sciences* 104.39, pp. 15181–15187. DOI: 10.1073/pnas.0702288104.

– (2009). "A General Framework for Analyzing Sustainability of Social-Ecological Systems". In: *Science* 325.5939, pp. 419–422. DOI: 10.1126/science.1172133.

– (2015). *Governing the Commons*. Cambridge University Press. DOI: 10.1017/cbo9781316423936.

Perc, M., J. Gómez-Gardeñes, A. Szolnoki, L. M. Floría, and Y. Moreno (2013). "Evolutionary dynamics of group interactions on structured populations: a review". In: *Journal of The Royal Society Interface* 10.80, pp. 20120997–20120997. DOI: 10.1098/rsif.2012.0997.

Perc, M. and A. Szolnoki (2010). "Coevolutionary games—A mini review". In: *Biosystems* 99.2, pp. 109–125. DOI: 10.1016/j.biosystems.2009.10.003.

Perc, M., J. J. Jordan, D. G. Rand, Z. Wang, S. Boccaletti, and A. Szolnoki (2017). "Statistical physics of human cooperation". In: *Physics Reports* 687, pp. 1–51. DOI: 10.1016/j.physrep.2017.05.004.

Perman, R., Y. Ma, J. McGilvray, and M. Common (2003). *Natural resource and environmental economics*. Pearson Education.

Pérolat, J., J. Z. Leibo, V. Zambaldi, C. Beattie, K. Tuyls, and T. Graepel (2017). "A multi-agent reinforcement learning model of common-pool resource appropriation". In: *Advances in Neural Information Processing Systems 30*, pp. 3643–3652.

Petschel-Held, G., H. J. Schellnhuber, T. Bruckner, F. L. Tóth, and K. Hasselmann (1999). "The tolerable windows approach: theoretical and methodological foundations". In: *Climatic Change* 41.3-4, pp. 303–331. DOI: 10.1023/a:1005487123751.

Pezzey, J. C. V. (1992). "Sustainable development concepts". In: *World Bank Environment Paper* 1.1, p. 45.

– (1997). "Sustainability Constraints versus "Optimality" versus Intertemporal Concern, and Axioms versus Data". In: *Land Economics* 73.4, p. 448. DOI: 10.2307/3147239.

– (2004). "One-sided sustainability tests with amenities, and changes in technology, trade and population". In: *Journal of Environmental Economics and Management* 48.1, pp. 613–631. DOI: 10.1016/j.jeem.2003.10.002.

Polasky, S., S. R. Carpenter, C. Folke, and B. Keeler (2011a). "Decision-making under great uncertainty: environmental management in an era of global change". In: *Trends in Ecology & Evolution* 26.8, pp. 398–404. DOI: 10.1016/j.tree.2011.04.007.

Polasky, S., A. De Zeeuw, and F. Wagener (2011b). "Optimal management with potential regime shifts". In: *Journal of Environmental Economics and Management* 62.2, pp. 229–240. DOI: 10.1016/j.jeem.2010.09.004.

Puterman, M. L. (2005). *Markov Decision Processes*. John Wiley & Sons, Inc. DOI: 10.1002/9780470316887.

Raffensperger, C. and J. A. Tickner (1999). *Protecting Public Health and the Environment: Implementing the Precautionary Principle*. Island Press: Wahington, DC / Covelo, California.

Raworth, K. (2012). "A safe and just space for humanity: can we live within the doughnut". In: *Oxfam Policy and Practice: Climate Change and Resilience* 8.1, pp. 1–16.

– (2017). "A Doughnut for the Anthropocene: humanity's compass in the 21st century". In: *The Lancet Planetary Health* 1.2, e48–e49. DOI: 10.1016/s2542-5196(17)30028-1.

Realpe-Gomez, J., B. Szczesny, L. Dall'Asta, and T. Galla (2012). "Fixation and escape times in stochastic game learning". In: *Journal of Statistical Mechanics: Theory and Experiment* 2012.10, P10022. DOI: 10.1088/1742-5468/2012/10/p10022.

Rocha, J., J. Yletyinen, R. Biggs, T. Blenckner, and G. Peterson (2014). "Marine regime shifts: drivers and impacts on ecosystems services". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 370.1659, pp. 20130273–20130273. DOI: 10.1098/rstb.2013.0273.

Rockström, J. and M. Klum (2012). *The Human Quest: Prospering within Planetary Boundaries*. Stockholm, Sweden: Langenskiölds.

Rockström, J., W. Steffen, K. Noone, A. Persson, F. S. Chapin, E. F. Lambin, T. M. Lenton, M. Scheffer, C. Folke, H. J. Schellnhuber, B. Nykvist, C. A. de Wit, T. Hughes, S. van der Leeuw, H. Rodhe, S. Sorlin, P. K. Snyder, R. Costanza, U. Svedin, M. Falkenmark, L. Karlberg, R. W. Corell, V. J. Fabry, J. Hansen, B. Walker, D. Liverman, K. Richardson, P. Crutzen, and J. A. Foley (2009a). "A safe operating space for humanity". In: *Nature* 461.7263, pp. 472–475. DOI: 10.1038/461472a.

Rockström, J., W. Steffen, K. Noone, Å. Persson, F. S. Chapin, III, E. F. Lambin, T. M. Lenton, M. Scheffer, C. Folke, H. J. Schellnhuber, B. Nykvist, C. A. de Wit, T. Hughes, S. van der Leeuw, H. Rodhe, S. Sörlin, P. K. Snyder, R. Costanza, U. Svedin, M. Falkenmark, L. Karlberg, R. W. Corell, V. J. Fabry, J. Hansen, B. Walker, D. Liverman, K. Richardson, P. Crutzen, and J. A. Foley (2009b). "Planetary Boundaries: Exploring the Safe Operating Space for Humanity". In: *Ecology and Society* 14.2, p. 32. DOI: 10.5751/es-03180-140232.

Rockström, J., O. Gaffney, J. Rogelj, M. Meinshausen, N. Nakicenovic, and H. J. Schellnhuber (2017). "A roadmap for rapid decarbonization". In: *Science* 355.6331, pp. 1269–1271. DOI: 10.1126/science.aah3443.

Rodriguez, A. A., O. Cifdaloz, J. M. Anderies, M. A. Janssen, and J. Dickeson (2010). "Confronting Management Challenges in Highly Uncertain Natural Resource Systems: a Robustness-Vulnerability Trade-off Approach". In: *Environmental Modeling & Assessment* 16.1, pp. 15–36. DOI: 10.1007/s10666-010-9229-z.

Rosa, H. (2013). *Social Acceleration*. Columbia University Press. DOI: 10.7312/rosa14834.

Roth, A. E. and I. Erev (1995). "Learning in extensive-form games: Experimental data and simple dynamic models in the intermediate term". In: *Games and Economic Behavior* 8.1, pp. 164–212. DOI: 10.1016/s0899-8256(05)80020-x.

Rougé, C., J.-D. Mathias, and G. Deffuant (2013). "Extending the viability theory framework of resilience to uncertain dynamics, and application to lake eutrophication". In: *Ecological Indicators* 29, pp. 420–433. DOI: 10.1016/j.ecolind.2012.12.032.

Sachs, L. (1984). *Applied Statistics*. Springer New York. DOI: 10.1007/978-1-4612-5246-7.

San Miguel, M., V. M. Eguiluz, R. Toral, and K. Klemm (2005). "Binary and Multivariate Stochastic Models of Consensus Formation". In: *Computing in Science and Engineering* 7.6, pp. 67–73. DOI: 10.1109/mcse.2005.114.

Sánchez, A. (2018). "Physics of human cooperation: experimental evidence and theoretical models". In: *Journal of Statistical Mechanics: Theory and Experiment* 2018.2, p. 024001. DOI: 10.1088/1742-5468/aaa388.

Sanders, J. B., T. Galla, and J. L. Shapiro (2012). "Effects of noise on convergent game-learning dynamics". In: *Journal of Physics A: Mathematical and Theoretical* 45.10, p. 105001. DOI: 10.1088/1751-8113/45/10/105001.

Sandri, M. (1996). "Numerical calculation of Lyapunov exponents". In: *The Mathematica Journal* 6.3, pp. 78–84.

Sato, Y. and J. P. Crutchfield (2003). "Coupled replicator equations for the dynamics of learning in multiagent systems". In: *Physical Review E* 67.1. DOI: 10.1103/physreve.67.015206.

Sato, Y., E. Akiyama, and J. D. Farmer (2002). "Chaos in learning a simple two-person game". In: *Proceedings of the National Academy of Sciences* 99.7, pp. 4748–4751. DOI: 10.1073/pnas.032086299.

Sato, Y., E. Akiyama, and J. P. Crutchfield (2005). "Stability and diversity in collective adaptation". In: *Physica D: Nonlinear Phenomena* 210.1-2, pp. 21–57. DOI: 10.1016/j.physd.2005.06.031.

Sayama, H., I. Pestov, J. Schmidt, B. J. Bush, C. Wong, J. Yamanoi, and T. Gross (2013). "Modeling complex systems with adaptive networks". In: *Computers & Mathematics with Applications* 65.10, pp. 1645–1664. DOI: 10.1016/j.camwa.2012.12.005.

Scheffer, M., S. Carpenter, J. A. Foley, C. Folke, and B. Walker (2001). "Catastrophic shifts in ecosystems". In: *Nature* 413.6856, pp. 591–596. DOI: 10.1038/35098000.

Schellnhuber, H. J. (1998). "Discourse: Earth System Analysis - The Scope of the Challenge". In: *Earth System Analysis*. Springer Berlin Heidelberg, pp. 3–195. DOI: 10.1007/978-3-642-52354-0_1.

– (1999). "Earth system analysis and the second Copernican revolution". In: *Nature* 402.6761supp, pp. C19–C23. DOI: 10.1038/35011515.

– (2009). "Tipping elements in the Earth System". In: *Proceedings of the National Academy of Sciences* 106.49, pp. 20561–20563. DOI: 10.1073/pnas.0911106106.

Schellnhuber, H. J., S. Rahmstorf, and R. Winkelmann (2016). "Why the right climate target was agreed in Paris". In: *Nature Climate Change* 6.7, pp. 649–653. DOI: 10.1038/nclimate3013.

Schill, C., T. Lindahl, and A.-S. Crépin (2015). "Collective action and the risk of ecosystem regime shifts: insights from a laboratory experiment". In: *Ecology and Society* 20.1. DOI: 10.5751/es-07318-200148.

Schleussner, C.-F., J. F. Donges, D. A. Engemann, and A. Levermann (2016). "Clustered marginalization of minorities during social transitions induced by co-evolution of behaviour and network structure". In: *Scientific Reports* 6.1. DOI: 10.1038/srep30790.

Schlüter, M., A. Baeza, G. Dressler, K. Frank, J. Groeneveld, W. Jager, M. A. Janssen, R. R. McAllister, B. Müller, K. Orach, N. Schwarz, and N. Wijermans (2017). "A framework for mapping and comparing behavioural theories in models of social-ecological systems". In: *Ecological Economics* 131, pp. 21–35. DOI: 10.1016/j.ecolecon.2016.08.008.

Shah, A. (2012). "Psychological and Neuroscientific Connections with Reinforcement Learning". In: *Reinforcement Learning*. Springer Berlin Heidelberg, pp. 507–537. DOI: 10.1007/978-3-642-27645-3_16.

Shapley, L. S. (1953). "Stochastic Games". In: *Proceedings of the National Academy of Sciences* 39.10, pp. 1095–1100. DOI: 10.1073/pnas.39.10.1095.

Shoham, Y. and K. Leyton-Brown (2008). *Multiagent systems: Algorithmic, game-theoretic, and logical foundations*. Cambridge University Press. DOI: 10.1017/cbo9780511811654.

Sinatra, R., P. Deville, M. Szell, D. Wang, A.-L. Barabási, R. Sinatra, P. Deville, M. Szell, D. Wang, and A.-L. Barabási (2015). "A century of physics". In: *Nature Physics* 11.10, pp. 791–796. DOI: 10.1038/nphys3494.

Smolla, M., R. T. Gilman, T. Galla, and S. Shultz (2015). "Competition for resources can explain patterns of social and individual learning in nature". In: *Proceedings of the Royal Society B: Biological Sciences* 282.1815, p. 20151405. DOI: 10.1098/rspb.2015.1405.

Snijders, T. A., G. G. Van de Bunt, and C. E. Steglich (2010). "Introduction to stochastic actor-based models for network dynamics". In: *Social Networks* 32.1, pp. 44–60. DOI: 10.1016/j.socnet.2009.02.004.

Spaan, M. T. J. (2012). "Partially Observable Markov Decision Processes". In: *Reinforcement Learning*. Springer Berlin Heidelberg, pp. 387–414. DOI: 10.1007/978-3-642-27645-3_12.

Stauffer, D. (2012). "A Biased Review of Sociophysics". In: *Journal of Statistical Physics* 151.1-2, pp. 9–20. DOI: 10.1007/s10955-012-0604-9.

Steffen, W., R. Sanderson, P. Tyson, J. Jäger, P. Matson, B. Moore III, F. Oldfield, K. Richardson, H. Schellnhuber, B. Turner, and R. Wasson (2005). *Global Change and the Earth System: A Planet Under Pressure*. Springer Berlin Heidelberg. DOI: 10.1007/b137870.

Steffen, W., K. Richardson, J. Rockström, S. E. Cornell, I. Fetzer, E. M. Bennett, R. Biggs, S. R. Carpenter, W. de Vries, C. A. de Wit, C. Folke, D. Gerten, J. Heinke, G. M. Mace, L. M. Persson, V. Ramanathan, B. Reyers, and S. Sorlin (2015a). "Planetary boundaries: Guiding human development on a changing planet". In: *Science* 347.6223, pp. 1259855–1259855. DOI: 10.1126/science.1259855.

Steffen, W., P. J. Crutzen, and J. R. McNeill (2007). "The Anthropocene: Are Humans Now Overwhelming the Great Forces of Nature". In: *AMBIO: A Journal of the Human Environment* 36.8, pp. 614–621. DOI: 10.1579/0044-7447(2007)36[614:TAAHNO]2.0.CO;2.

Steffen, W., W. Broadgate, L. Deutsch, O. Gaffney, and C. Ludwig (2015b). "The trajectory of the Anthropocene: The Great Acceleration". In: *The Anthropocene Review* 2.1, pp. 81–98. DOI: 10.1177/2053019614564785.

Steffen, W., J. Rockström, K. Richardson, T. M. Lenton, C. Folke, D. Liverman, C. P. Summerhayes, A. D. Barnosky, S. E. Cornell, M. Crucifix, J. F. Donges, I. Fetzer, S. J. Lade, M. Scheffer, R. Winkelmann, and H. J. Schellnhuber (2018). "Trajectories of the Earth System in the Anthropocene". In: *Proceedings of the National Academy of Sciences* 115.33, pp. 8252–8259. DOI: 10.1073/pnas.1810141115.

Stern, N. (2008). "The Economics of Climate Change". In: *American Economic Review* 98.2, pp. 1–37. DOI: 10.1257/aer.98.2.1.

Sugiarto, H. S., N. N. Chung, C. H. Lai, and L. Y. Chew (2015). "Socioecological regime shifts in the setting of complex social interactions". In: *Physical Review E* 91.6, p. 062804. DOI: 10.1103/physreve.91.062804.

Sutton, R. S. and A. G. Barto (1998). *Reinforcement learning: An introduction*. MIT Press.

Szolnoki, A., M. Perc, and G. Szabó (2009). "Phase diagrams for three-strategy evolutionary prisoner's dilemma games on regular graphs". In: *Physical Review E* 80.5. DOI: 10.1103/physreve.80.056104.

Tavoni, A. and S. Levin (2014). "Managing the climate commons at the nexus of ecology, behaviour and economics". In: *Nature Climate Change* 4.12, pp. 1057–1063. DOI: 10.1038/nclimate2375.

Tavoni, A., A. Dannenberg, G. Kallis, and A. Löschel (2011). "Inequality, communication, and the avoidance of disastrous climate change in a public goods game". In: *Proceedings of the National Academy of Sciences* 108.29, pp. 11825–11829. DOI: 10.1073/pnas.1102493108.

Tavoni, A., M. Schlüter, and S. Levin (2012). "The survival of the conformist: Social pressure and renewable resource management". In: *Journal of Theoretical Biology* 299, pp. 152–161. DOI: 10.1016/j.jtbi.2011.07.003.

Traulsen, A., D. Semmann, R. D. Sommerfeld, H.-J. Krambeck, and M. Milinski (2010). "Human strategy updating in evolutionary games". In: *Proceedings of the National Academy of Sciences* 107.7, pp. 2962–2966. DOI: 10.1073/pnas.0912515107.

Tuyls, K. and A. Nowé (2005). "Evolutionary game theory and multi-agent reinforcement learning". In: *The Knowledge Engineering Review* 20.01, pp. 63–90. DOI: 10.1017/s026988890500041x.

Tuyls, K. and S. Parsons (2007). "What evolutionary game theory tells us about multiagent learning". In: *Artificial Intelligence* 171.7, pp. 406–416. DOI: 10.1016/j.artint.2007.01.004.

Tuyls, K., K. Verbeeck, and T. Lenaerts (2003). "A Selection-Mutation Model for Q-learning in Multi-agent Systems". In: *Proceedings of the Second International Joint Conference on Autonomous Agents and Multiagent Systems*. AAMAS 2003, pp. 693–700. DOI: 10.1145/860575.860687.

Tuyls, K., P. J. Hoen, and B. Vanschoenwinkel (2006). "An Evolutionary Dynamical Analysis of Multi-Agent Learning in Iterated Games". In: *Autonomous Agents and Multi-Agent Systems* 12.1, pp. 115–153. DOI: 10.1007/s10458-005-3783-9.

UN General Assembly (1948). *Universal Declaration of Human Rights*. Declaration. United Nations.

– (2015). *Transforming our world: The 2030 agenda for sustainable development - A/RES/70/1*. Resolution. United Nations.

van Vuuren, D. P., P. L. Lucas, T. Häyhä, S. E. Cornell, and M. Stafford-Smith (2016). "Horses for courses: analytical tools to explore planetary boundaries". In: *Earth System Dynamics* 7.1, pp. 267–279. DOI: 10.5194/esd-7-267-2016.

van Vuuren, D. P., L. B. Bayer, C. Chuwah, L. Ganzeveld, W. Hazeleger, B. van den Hurk, T. van Noije, B. O'Neill, and B. J. Strengers (2012). "A comprehensive view on climate change: coupling of earth system and integrated assessment models". In: *Environmental Research Letters* 7.2, p. 024012. DOI: 10.1088/1748-9326/7/2/024012.

Vasconcelos, V. V., F. C. Santos, J. M. Pacheco, and S. A. Levin (2014). "Climate policies under wealth inequality". In: *Proceedings of the National Academy of Sciences* 111.6, pp. 2212–2216. DOI: 10.1073/pnas.1323479111.

Verburg, P. H., J. A. Dearing, J. G. Dyke, S. van der Leeuw, S. Seitzinger, W. Steffen, and J. Syvitski (2016). "Methods and approaches to modelling the Anthropocene". In: *Global Environmental Change* 39, pp. 328–340. DOI: 10.1016/j.gloenvcha.2015.08.007.

Verendel, V., D. J. Johansson, and K. Lindgren (2015). "Strategic reasoning and bargaining in catastrophic climate change games". In: *Nature Climate Change* 6.3, pp. 265–268. DOI: 10.1038/nclimate2849.

von der Osten, F. B. (2017). "Intelligent decision-making in coupled socio-ecological systems". PhD thesis. The University of Melbourne.

Vrancx, P., K. Tuyls, and R. Westra (2008). "Switching dynamics of multi-agent learning". In: *Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent systems*. AAMAS 2008, pp. 307–313.

Walker, J. M. and R. Gardner (1992). "Probabilistic Destruction of Common-pool Resources: Experimental Evidence". In: *The Economic Journal* 102.414, p. 1149. DOI: 10.2307/2234382.

Wang, Z., L. Wang, A. Szolnoki, and M. Perc (2015). "Evolutionary games on multilayer networks: a colloquium". In: *The European Physical Journal B* 88.5. DOI: 10.1140/epjb/e2015-60270-7.

WCED (1987). *World Commission on Environment and Development - Our Common Future*. Report. United Nations.

Weyant, J. (2014). "Integrated assessment of climate change: state of the literature". In: *Journal of Benefit-Cost Analysis* 5.03, pp. 377–409. DOI: 10.1515/jbca-2014-9002.

Wiedermann, M., J. F. Donges, J. Heitzig, W. Lucht, and J. Kurths (2015). "Macroscopic description of complex adaptive networks coevolving with dynamic node states". In: *Physical Review E* 91.5. DOI: 10.1103/physreve.91.052801.

Wiener, N. (1961). *Cybernetics or Control and Communication in the Animal and the Machine*. MIT Press. DOI: 10.1037/13140-000.

Wiering, M and M. van Otterlo (2012). *Reinforcement Learning: State-of-the-Art*. Springer Berlin Heidelberg. DOI: 10.1007/978-3-642-27645-3.

Woodward, R. T. and D. Tomberlin (2014). "Practical Precautionary Resource Management Using Robust Optimization". In: *Environmental Management* 54.4, pp. 828–839. DOI: 10.1007/s00267-014-0348-1.

Worm, B., R. Hilborn, J. K. Baum, T. A. Branch, J. S. Collie, C. Costello, M. J. Fogarty, E. A. Fulton, J. A. Hutchings, S. Jennings, O. P. Jensen, H. K. Lotze,

P. M. Mace, T. R. McClanahan, C. Minto, S. R. Palumbi, A. M. Parma, D. Ricard, A. A. Rosenberg, R. Watson, and D. Zeller (2009). "Rebuilding Global Fisheries". In: *Science* 325.5940, pp. 578–585. DOI: 10.1126/science.1173146.

Zanette, D. H. and S. Gil (2006). "Opinion spreading and agent segregation on evolving networks". In: *Physica D: Nonlinear Phenomena* 224.1-2, pp. 156–165. DOI: 10.1016/j.physd.2006.09.010.

Zhou, K. and J. C. Doyle (1998). *Essentials of robust control*. Upper Saddle River, N.J. : Prentice Hall.

Selbständigkeitserklärung

Ich erkläre, dass ich die Dissertation selbständig und nur unter Verwendung der von mir gemäß §7 Abs. 3 der Promotionsordnung der Mathematisch-Naturwissenschaftlichen Fakultät, veröffentlicht im Amtlichen Mitteilungsblatt der Humboldt-Universität zu Berlin Nr. 126/2014 am 18.11.2014 angegebenen Hilfsmittel angefertigt habe.

Berlin, den 17. Dezember 2018 Wolfram Martin Barfuß